

國立臺灣大學生命科學院基因體與系統生物學學位學程

碩士論文

Genome and Systems Biology Degree Program

College of Life Science

National Taiwan University

Master Thesis



建構線上DNA變異位點插補伺服器

Development of an online system for DNA imputation

沈映成

Ying-Cheng Shen

指導教授：莊曜宇 博士

Advisor: Eric Y. Chuang, Sc.D.

中華民國 110 年 7 月

July, 2021

誌謝



時光飛逝，日月如梭，兩年的碩士生活就這樣結束了，能夠順利完成這篇論文實在有許多需要感謝的人。首先，要感謝指導教授莊曜宇老師對實驗室的付出與指導，老師建立起來的環境讓我們能夠放心的進行學習，雖然老師平時相當繁忙，但是在閒暇之餘老師還是會撥空來關心我們的狀況，非常謝謝老師對我的指導。也感謝實驗室中的各位老師，謝謝盧子彬老師在我的研究過程中指點方向，謝謝賴亮全老師以及蔡孟勳老師總是在實驗室會議中提點出研究中的問題，讓我能夠進行改善。也非常感謝郭錦輯主任能夠撥空來作為口試委員為我提供寶貴的建議，使得論文能夠更加完善。

在實驗室的生活我認識了許多人，實驗室的學長學姊們都相當照顧我，佳興學長、源懋學長在我還剛進實驗室的時候帶領我尋找方向以及熟悉環境，在研究中有困難的時候學長們也給了我很大的幫助。韋霓學姊時常邀請我們一起吃飯讓我更加認識實驗室的同學以及學長姐們。當我對碩士新生活以及研究內容感到迷惘時，晏均學長、育昇學長、惠政學姊、凱元學長也會跟我們分享他們過去的經驗給我。在課業上，要謝謝航凱學長、人豪學長、以及所上各位同學的幫助，讓我能順利學習並通過這兩年的充實課程。也要特別感謝 amrita 學姊細心教導我的研究內容以及論文，非常謝謝這些學長姐的在我學習過程中的幫助。

最後，要感謝我的同儕、學弟妹和朋友們，謝謝嘉峻、玠妏、晉懷、昱衡以及兆榮在學習過程中陪我一同努力，謝謝 SASA、哈破、以及群裡的朋友們為我帶來歡樂，因為有你們給予的幫助與陪伴，我才能夠度過這兩年的重重困難，並完成這篇論文，這兩年得到的一切是無比珍貴的回憶，我會將其銘記在心中。祝福大家一切平安、順利！

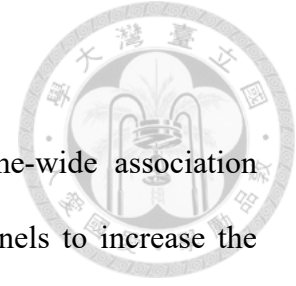
沈映成 謹誌於
台灣大學基因體與系統生物學學位學程
民國一一〇年八月

中文摘要



基因型插補(Genotype imputation),是在進行全基因組關聯研究(Genome wide association study ,GWAS)之前的重要步驟。它能透過龐大的參考序列資料庫進行預測並填補缺失的基因型以增加樣品的 SNP 密度和 GWAS 的分析資料性。然而,整個基因型插補過程包括一系列複雜的插補前以及插補後步驟,運算過程需要耗費龐大的運算資源量,並且也需要生物資訊學專業知識。因此,我們建立了一個對於使用者方便的網頁插補伺服器服務,名為 Multi-racial Imputation System (MI-system),該服務分別使用生物資訊學家常用的 pre-phasing 軟體 SHAPEIT 和 imputation 軟體 IMPUTE2 進行運算。對於所使用的參考序列資料庫,該服務首次包括了 Taiwan biobank (TWB) 序列資料庫,並根據使用者需求為其提供 1000 Genome Phase III 和 TWB 以及 Hapmap3 序列資料庫可進行選擇,也添加了 IMPUTE2 特有的兩種 merge reference imputation 功能來增強插補的結果。該服務進一步提供了彈性的質量控制選項,並讓使用者能從多個選項中自行選擇所要篩選的次要等位基因頻率(Minor allele frequency)閾值、需要過濾的基因型及樣本缺失率以及 Hardy-Weinberg 平衡的閾值。為了增加使用者的便利性,該服務還提供了一些實用功能,例如 (i) 分割全基因組 SNP 資料,(ii) 基因組座標軸轉換 (grch37 和 grch38),以及 (iii) 透過使用者上傳的基因型資料建立定制建構參考序列資料庫。使用者只需要簡單的幾個步驟即可執行實用程式功能並快速獲得高通量的 SNP 插補資料。並能夠將結果轉換成與流行的 GWAS 分析工具(例如 PLINK, SNPTest 或 R)兼容的格式進行下載,以方便進行後續分析。

Abstract



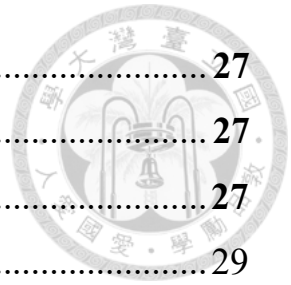
Genotype imputation is an important process before genome-wide association studies (GWAS). It predicts missing genotypes from reference panels to increase the sample SNP density and the power of the GWAS. However, the process encompasses a series of extensive pre and post imputation procedures, is computationally expensive, and requires expertise in bioinformatics. Therefore, we have developed a user-friendly, web-based imputation service, Multi-racial Imputation System (MI-system), that utilizes popular pre-phasing and imputation softwares, SHAPEIT2 and IMPUTE2, respectively. For the reference panels, the server includes the Taiwan biobank (TWB) panel for the first time. It offers users to choose from 1000 Genome phase III, TWB and Hapmap3 panels, as reference genomes. Furthermore, the users can choose the IMPUTE2 specific function “merge reference”, for merging multiple reference panels to conduct imputation. The server, also provides flexible quality control options and allows users to choose thresholds for parameters such as minor allele frequency, missing (SNP level and individual level) genotyping rates, and Hardy-Weinberg equilibrium. For user’s convenience several additional utility functions such as (i) splitting whole genome SNP data, (ii) conversion of genome builds (grch37 and grch38), and (iii) build customized reference panels from user uploaded genotype data, are offered. The users can obtain high-throughput imputed data and access utility functions through few easy and simple clicks. The results can also be downloaded in formats that are compatible with popular GWAS tools such as PLINK, SNPTest, or R to further downstream analysis.

Contents



誌謝	ii
中文摘要	iii
Abstract	iv
Contents	v
List of figures	vii
List of tables	ix
Chapter 1 Introduction	1
1.1 Single Nucleotide Polymorphism (SNP).....	4
1.2 Genome Wide Association Studies	5
1.3 SNP genotyping: Microarrays and Next generation sequencing..	6
1.4 Genotype imputation	7
1.4.1. The introduction of genotype imputation	7
1.4.2 Genotype imputation of Rare variants	8
1.4.3 The steps of genotype imputation	9
1.5 Reference panel in genotype imputation	12
1.6 Imputation Server	14
1.7 Specific aim	15
Chapter 2 Materials and methods	17
2.1 System implementation.....	17
2.2 MI-System Reference Panels.....	18
2.3 MI-System: Services.....	19
2.3.1 Service 1: Imputation	20
2.3.2 Service 2: Create reference panel.....	24
2.3.3 Service 3: Split Chromosome.....	25
2.3.4 Service 4: Liftover.....	25
Chapter 3 Results	27

3.1 Web-interface:	27
3.2 Public reference panels	27
3.3 Service: Imputation	27
3.3.1 Improve the cost of imputation time	29
3.3.2 Comparison of MI-System with Michigan Imputation server... 29	
3.3.3 Merge reference panels	32
3.3.4 Validate the imputation accuracy.....	33
3.4 Service: Split chromosome	34
3.5 Service: Liftover	35
3.6 Create reference	36
Chapter 4 Discussion	36
Chapter 5 Conclusions	39
Chapter 6 References	40
Appendix	53
Supplementary figure	79

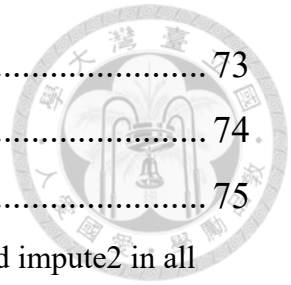


List of figures



Figure 1: The architectural design of the MI-System	53
Figure 2: The overview workflow of the MI-System	54
Figure 3: The workflow of the imputation function in MI-System	55
Figure 4: The homepage and the email entering window of the MI-System.....	56
Figure 5: The Venn Diagram of the common reference panels and TWB.....	57
Figure 6: The imputation page of the MI-System.....	58
Figure 7: The result page of the MI-System	59
Figure 8: The download page of the MI-System	60
Figure 9: The info score distribution plot of the imputation result.....	61
Figure 10: The rare SNPs distribution plot (rare variant) of the imputation result.....	61
Figure 11: The common SNPs distribution plot (common variant) of the imputation result.....	62
Figure 12: The parallel computation design of the imputation.....	63
Figure 13: The common SNPs distribution plots from MI-System and Michigan Imputation server (EAS group)	64
Figure 14: The common SNPs distribution plots from MI-System and Michigan Imputation server (EUR group).....	65
Figure 15: The rare SNPs distribution plots from MI-System and Michigan Imputation server (EAS group)	66
Figure 16: The rare SNPs distribution plots from MI-System and Michigan Imputation server (EAS group)	67
Figure 17: Rsq comparison plots of accuracy between Minimac4 and impute2.....	68
Figure 18: Comparison of common SNPs imputed for 4 different reference panels ...	69
Figure 19: Comparison of rare SNPs imputed for 4 different reference panels.....	70
Figure 20: Accuracy comparison between 4 reference panels (Total SNPs)	71
Figure 21: Accuracy comparison between 4 reference panels (Rare SNPs)	72

Figure 22: The split chromosome page of the MI-System.....	73
Figure 23: The liftover page of the MI-System	74
Figure 24: The Create reference page of the MI-System	75
Figure S1: Rsq comparison plots of accuracy between Minimac4 and impute2 in all chromosomes (EAS group)	79
Figure S2: Rsq comparison plots of accuracy between Minimac4 and impute2 in all chromosomes (EUR group)	80
Figure S3: Accuracy comparison between 4 reference panels in all chromosomes (Total SNPs)	81
Figure S4: Accuracy comparison between 4 reference panels in all chromosomes (Rare SNPs)	82

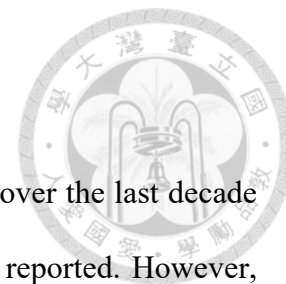


List of tables



Table 1: The validation of imputation accuracy	76
Table 2: The imputation cost time of the IMPUTE2	77
Table 3: The comparison table between several imputation servers and MI-System ..	78

Chapter 1 Introduction



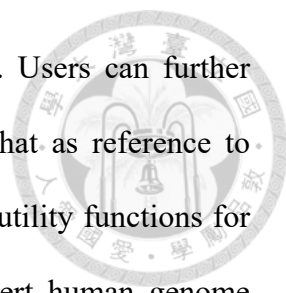
Complex disease association studies have been revolutionized over the last decade and a half, with nearly 50,000 genome wide significant loci being reported. However, GWAS have had been under controversy, because of the “missing heritability” issue. Putative loci that are identified by GWAS explain only a fraction of the heritability of complex diseases [1], thereby leading to research endeavors that can account for the unexplained heritability. One of the reasons behind this is that SNP arrays are constructed from variants that are selected based on linkage disequilibrium (LD). Hence, the discovery of putative variants basically rests on the pair-wise LD of common variants with causal ones [2]. However, variants with lower frequencies (minor allele frequency (MAF) <0.01) in populations (rare variants) have lower LDs with causal ones and thereby goes undetected, in spite of them playing a significant part in disease etiology [3]. Genotype imputation is one of the popular directions that have been employed to address this issue as it utilizes whole genome sequencing reference panels from public domains to estimate un-typed variants (single nucleotide polymorphism (SNPs)) into a low dense genotype panel [4]. With the increasing coverage of such panels, rare variants get successfully detected with increased accuracy, because imputation allows better representation of these variants in the study population. Imputation is a powerful approach that leads to (i) better identification of causal rare variants, (ii) facilitation of fine mapping studies and identification of specific causal loci from GWAS susceptible regions [5], (iii) converging multiple study datasets for meta-analysis by filling up missing SNPs for each of the datasets [6], and (iv) increasing the power of the studies [7]. Genotype imputation has been quite successful for common and low-frequency variants, utilizing multi ethnic reference panels such as

1000 Genomes (1KG) [8], international Hapmap project [9], Haplotype reference consortium (HRC) [10] and TopMED [11].

Recently, various studies have reported that ethnically heterogeneous reference panels, have failed to provide rare variant imputation with higher accuracy [12]. A prior study on cystic fibrosis (CF) pointed out that public datasets such as 1KG or HRC panels lack the causal CF specific genomic regions [13]. Hence CF-associated haplotypes would eventually be omitted from GWAS as they would fail to get imputed. Including matched ethnic specific panels in addition to multi ethnic panels, have been shown to provide better imputation accuracy, specifically for rarer variations [14, 15]. Therefore, based on the study goal, it is crucial that an appropriate reference panel is employed to conduct imputation towards more meaningful and accurate findings.

Merging reference panels involves multi-step procedures. First, successive forward and backward imputations are required between study specific and multi ethnic reference panels, before merging them [16]. Once the merged panels are created a series of procedures including, quality control, pre-phasing and imputation needs to be conducted to finally obtain a suitable GWAS panel. The datasets involved, in the process are massive. Conducting, all of the above successfully, requires bioinformatics expertise and powerful computation resources. As the processes involves managing high dimensional datasets, high end processors with provision for parallel computing is a pre-requisite for simultaneous computation towards significant reduction of time for imputation and obtaining accurate results. Such requirements pose a bottleneck, especially for non- bioinformatics personnel.

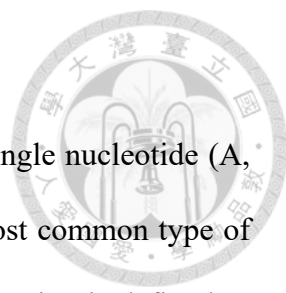
Therefore, in this study we present a public imputation and analysis platform, Multi-racial Imputation System (MI-System), to allow users with access to a high speed



and efficient platform to seamlessly conduct genotype imputation. Users can further create, reference panels using their customized datasets and use that as reference to conduct imputation. To ensure convenience, the system also offers utility functions for the users to split the whole genome into chromosomes and convert human genome builds between hg37 and hg38 with easy clicks. The goal is to eliminate the need for bioinformatics expertise, which is otherwise a pre-requisite for obtaining highly accurate imputed datasets.

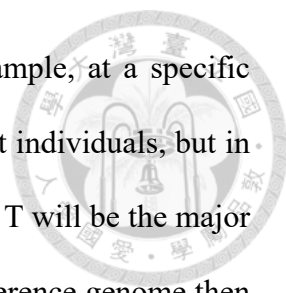
One of the most popular public imputation platforms is the Michigan Imputation Server, which has been making great contributions into the field of population genetics research [17]. However, their primary focus has been the European and Caucasian populations. The latest version of the Michigan server uses Eagle2 – Minimac4 [18, 19] to conduct pre-phasing and imputation, respectively. To the best of our knowledge, a large section of the imputation community, uses SHAPEIT2-IMPUTE2 [20, 21] as their choice of software for pre-phasing and imputation. For users such as these and more, MI-System, opens up the opportunity to expand their research boundaries by enabling them to conduct imputation, create and merge reference panels, using any target population of any ancestry, with SHAPEIT2-IMPUTE2. Furthermore, MI-system provides the users with a customized Taiwan Biobank (TWB) reference panel which is specifically created for people of Taiwanese-Chinese ancestry. We believe that MI-System would potentially have a significant contribution to the field of population genetics research, and take it few notches further, by working hand in hand with the existing servers.

1.1 Single Nucleotide Polymorphism (SNP)



Single nucleotide polymorphism (SNP), is the alteration of a single nucleotide (A, T, C, G) at a specific position in a DNA sequence. SNPs are the most common type of genetic variation that are observed among individuals. A genetic mutation is defined as any change in a DNA sequence away from normal, where a normal allele that is prevalent in the population is changed to a rare and abnormal variant. In contrast, a polymorphism is a DNA sequence variation in the population that is commonly observed. SNPs are detectable in >1% of the population whereas DNA mutations found in <1 % of the population [22]. In the human genome, one SNP appears every 100 to 300 bases, and 90% of the differences in genes expression had been found to be related to SNPs [23]. They may be responsible for the diversity among different ethnic populations in the most common familial traits such as eye color and differences in height. Other traits such as differences in drug response between individuals, diseases such as, obesity, diabetes, autoimmune diseases, psychiatric disorders, and cancer susceptibility [24-32], have been reported to be associated with SNPs

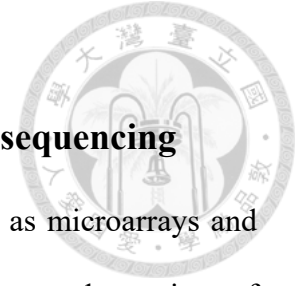
The occurrence frequency of a specific nucleotide at an SNP site is called allele frequency (AF). AF is calculated by dividing the number of times the allele of interest is observed in a population by the total number of copies of all the alleles at that particular genetic locus in the population. The reference allele simply refers to whether an allele matches an SNP-specific region in the reference genome (for humans: human reference genome: GRCh37 or GRCh38) and the alternate allele in contrast, refers to any base, other than the reference, that is found at that locus. The more frequently occurring allele is called the major allele, and the other is called the minor allele. The alternative allele is not necessarily the minor allele and it may, or may not, be linked to a phenotype.



There can be more than one alternative allele per variant. For example, at a specific SNP site in the human genome, the T nucleotide may appear in most individuals, but in a minority of individuals, the nucleotide in the site is a C. Therefore, T will be the major allele and C will be the minor allele. If T is present in the human reference genome then T will be the reference allele, otherwise the alternate allele. The frequency of the minor allele in the study population is called the minor allele frequency (MAF). According to the minor allele frequency of nucleotides at the SNP site, the SNP can be classified into common variant ($MAF \geq 5\%$), less common variant ($0.01 < MAF < 0.05$) and rare genetic variant ($MAF \leq 0.01$) [33].

1.2 Genome Wide Association Studies

Genome-wide association studies (GWAS) [34] are used to analyze and screen the relationship between SNPs and phenotypes in the human genome. It tests hundreds of thousands to millions of genetic variants across the individual genomes to identify genotype–phenotype associations. The analysis of GWAS requires the population to be divided into experimental group and control group according to the phenotype that they want to observe. Then, obtain the SNP data from the two populations and compare which SNPs are significantly associated with the experimental group. Further, study of the relationship between these SNP-related genes can provide more clues to the pathogenesis of complex diseases or the phenotype [35]. This allows pinpointing gene markers in addition to the traditional factors such as age, sex, family history, that may be associated with the disease of interest, through large-scale whole-genome analysis. Furthermore, the interplay of genetic markers and environmental factors can be considered to calculate the incidence of some diseases including cancer [36].



1.3 SNP genotyping: Microarrays and Next generation sequencing

SNP genotyping is mainly conducted using technologies such as microarrays and next generation sequencing (NGS) [37, 38]. SNP microarray protocol consists of coating primers with complementary nucleotides of the known SNP nucleotide as probes to a surface such as that of a silicon chip, glass beads or magnetic beads to build up a microarray chip. Once done, polymerase chain reaction (PCR) is used to amplify the individual DNA and break it into small sequences by enzyme. Finally, the sample DNA sequences are hybridized on the microarray, where the complementary nucleotides of the DNA get attached to each of the microarray probes on the chip. The fluorescence probes are then utilized to detect the signal and get the SNP sequences [39]. For NGS technology, the sample DNA is first broken down into small sequences using ultrasonic technologies and an adapter used to mark each sequence. Then, again PCR is used to amplify the small sequences, after which the dNTPs are added and the fluorescence tags are marked as start points of sequencing. The sequences are obtained by real-time detection of the fluorescent signals. Finally, software like SAMTOOLS are utilized to conduct variant calling [40]. Both methods have their respective advantages, SNP array is cheaper and more accurate, NGS can get more sequence information [41, 42]. On the other hand, SNP arrays from different manufacturers such as Illumina or Affymetrix cover different sets of SNP sites, which may lead to loss of information in further analysis. Depending on the sample quality, machine error rate, and sequencing coverage, SNP data obtained through NGS technology will also tend to lose SNP data. Such missing data renders incorrect inferences when GWAS is conducted [43, 44]. Therefore, genotype imputation is a cost-effective way that researchers often use to fill

out missing SNP data towards the improvement of the power of further analysis.



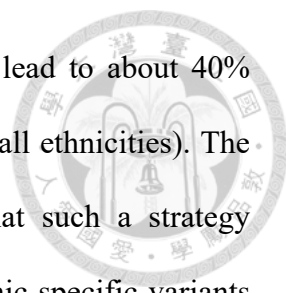
1.4 Genotype imputation

1.4.1. The introduction of genotype imputation

With the progression of precision medicine, variants (SNPs) have been proved to be more and more important. SNPs have been characterized in terms of their co-dominance, reproducibility, locus-specificity, and random genome-wide distribution, and a detailed analysis of SNPs can identify pathogenic ones in individual patients. This is partly because SNPs have considerable effects on protein function in coding sequences and gene expression in regulatory regions [45]. Hence, SNPs are the ideal candidates for genetics research, leading to functional characterization and identification of associated traits. However, current genotyping methods for SNP data often result in many missing data due to cost issues and the machine error, and the missing data prevents further analysis such as genome-wide association studies (GWAS) to obtain comprehensive information [46,47]. This is where genotype imputation comes in useful. Genotype imputation refers to the statistical inference of unobserved genotypes, using a denser reference panel, to fill out the missing SNP data and improve the power of association studies. The principle of genotype imputation is based on linkage disequilibrium (LD) that exists between the SNPs in close proximity. Genotypes are imputed using LD of flanking SNPs to the target SNP using the reference genome. LD is the non-random association of neighboring SNPs meaning they have a lower chance of crossing over, thus causing two SNPs to be in genetic linkage [48, 49]. By using the genotype imputation, the original $10^5\sim 10^6$ sites of SNP data can be filled up to the $10^7\sim 10^8$ high density SNP data and thus can improve the power of the GWAS.

1.4.2 Genotype imputation of Rare variants

Rare variants constitute the bulk of genetic variation in the human genome and are predicted to have larger phenotypic effects than common variants; however, it has been challenging to analyze such variants with adequate power in population-based studies due to their poor representation in the genotyping arrays that are typically used in GWASs [50]. The total number of loci that may contribute to a disease's prevalence is dependent on the disease incidence, the frequency of rare variants per locus, and their effect size (the genotype relative risk). For a disease with high heritability, with the increase of the number of contributing rare alleles in an individual, the relative risk rises steeply under a multiplicative model. However, if each of these variants explains most of the risk in just a few people, their effects will not explain enough of the variance in a total population. Therefore standard GWAS procedures would fail to detect them. Furthermore, they are scarcely tagged as single-nucleotide polymorphisms (SNPs) in genome-wide arrays, with the exception of family-based studies and studies with very large sample sizes. Moreover, most common disease common variant based GWASs exclude rare variants from their analysis in the early quality control steps. Some studies have shown that combining reference panels may increase the number and accuracy of imputed rare variants in comparison to when they are imputed using single reference panels [51]. Hence, a customized SNP panel along with other reference panels, such as 1000 Genomes Phase 3, could be combined to identify additional rare SNPs by imputation. Using a reference panel from multiple ethnic groups for SNPs that are not population-specific may still be inappropriate to provide accurate imputation. However, combining a global reference genome, such as 1000 Genomes [using only population specific samples, such as only East Asian (EAS) or only European (EUR)], with that of



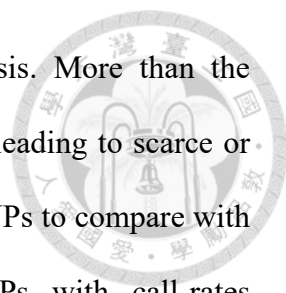
a respective ethnically specific one, to conduct imputation, could lead to about 40% more imputed variants than when using 1000 Genomes only (using all ethnicities). The improvement of imputation accuracy is attributed to the fact that such a strategy successfully captures the linkage disequilibrium patterns of the ethnic specific variants which other ethnic groups with different ancestral genetic backgrounds, might fail to capture. The ethnic composition is an important predictor of imputation accuracy. Many studies have validated the accuracy and reliability of imputation of rare variants, but the focus of most of these studies has been on populations of European descent [52]. Little research has been conducted in the area of rare variant imputation across asian populations, and none in Taiwanese populations. Therefore, utilizing population-based resources of Taiwanese ancestry, to construct reference panels would enable successful rare variant imputation for the Taiwanese population.

1.4.3 The steps of genotype imputation

The steps of the genotype imputation can be divided into three parts, quality control, pre-phasing and imputation. Each of the steps has a different function and special settings are required to be conducted for each of the steps using different softwares.

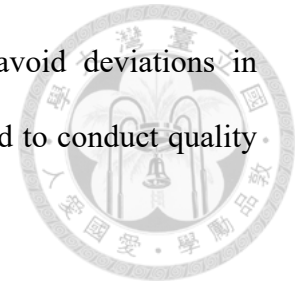
Quality control:

The first step is quality control, in which the primary purpose is to remove data that are of poor quality in the DNA genotype sample. It is of extreme importance to remove poor quality SNPs before proceeding for further analysis, as lesser data quality can lead to inaccurate imputations which can later produce false positives and false negatives from genome wide association analysis [53]. SNPs with low call rate and individuals



with high missing rate should be excluded before further analysis. More than the allowable missing-ness, leads to removal of many SNPs therefore, leading to scarce or no SNPs in the genotype data. Hence, there would be insufficient SNPs to compare with the reference, resulting in inaccurate imputation. Ideally, SNPs with call-rates >90%-95% in study subjects, and individuals with < 5-10% missing genotypes are retained for further analysis. Minor allele frequency (MAF) also affects the imputation accuracy [54]. Furthermore, when the minor allele frequency of a specific SNP site is too low, it means that the SNP is quite rare in the population, which makes it difficult to find the corresponding haplotype when comparing it with the reference panels, thus affecting the imputation accuracy. Studies have shown that higher GWAS chip density does have a positive effect on imputation quality. However, it is also argued that rarer SNPs are usually excluded as they may have a negative effect on the accuracy, due to poor genotype calling, and lack of linkage disequilibrium (LD) [55]. Hence careful thought is required in choosing thresholds for exclusion and inclusion of SNPs. However, the past decade has witnessed that these rare MAF SNPs are often associated with specific diseases [56], hence current research focus is to conduct GWAS using rare SNPs. Thus imputation strategies are undergoing modifications and improvisations to enhance the representation of rarer variants in populations, so that they can get included in GWAS studies and are found to have a greater effect size than what it was in the past. Besides that, the DNA genotype samples also are checked for Hardy-Weinberg Equilibrium (HWE). Suppose genotype or allele frequencies deviate significantly from HWE. In that case, it can indicate systematic errors in genotyping, unexpected population structure, presence of homologous regions in the genome, association with trait in case-control studies [57]. Since the last of those is least likely, so deviation from

HWE is an indicator that a marker should be discarded, and avoid deviations in subsequent imputation. Plink [58] is a popular tool that is often used to conduct quality control steps on genotype data.



Pre-phasing:

Pre-phasing is conducted on the study genotype data in concordance with the reference genome. Homo-sapiens (humans) have two copies of each chromosome (diploid), one of which is inherited from the father and the other from the mother. The technologies used for sequencing, generate short read sequences for individuals which are assembled into a single sequence. Phasing is a process that constitutes of distinguishing the paternal and maternally obtained chromosomal strands into homologous chromosomal pairs [59]. This requires complicated statistical computations, that uses the LD structure of the SNPs in the genotype and the reference genome to reconstruct haplotypes. This step is critical as it restores the functional consequences that was originally to be, in context to shared ancestry. While performing imputation, pre-phasing allows a one-time phasing where a phased genotype data can match to a selected strand of the reference panel, instead of matching two un-phased genotypes, therefore speeding up the consecutive imputation steps while reducing the cost of computation. The pre-phasing step is divided into three steps, alignment, strand check and phasing. The commonly used phasing tools are SHAPEIT and Eagle. For the alignment step, the study genotype data will be checked for its alignment with the reference panel, while marking the SNPs that are misaligned or are not aligned to the forward strand of the reference genome. The second step constitutes of removing SNPs from the study genotype data that failed the alignment and strand check step and hence marked in the first step. Finally, the third step consists of phasing the genotype data in

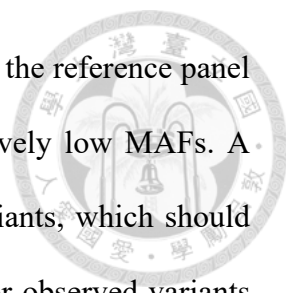
accordance with the reference panel.

Imputation:

For the imputation step, IMPUTE2 and minimac4 are the tools that are popularly employed. The step here will use the data from the pre-phasing result, and divide the imputed chromosome into several chunks, and start to calculate the probability of the nucleotides in each SNP site. Each of the above tools uses different algorithms to conduct imputation. In general, IMPUTE2's algorithm is more accurate, while the algorithm used in minimac4 is suitable for the improvement of the imputation speed. Based on the SNPs from user uploaded genotype data, the corresponding matched haplotype(s) get selected from the reference panel and the alleles from the haplotype region are utilized to infer the missing genotypes into the user panel. Absence of SNPs in each of the base-pair region of the study genotype data, fails to impute SNPs in missing sites. Therefore, each imputation chunk is flanked at either side by an internal 250kb buffer region, by IMPUTE2, to enhance the probability of type-2 SNPs in each chunk.

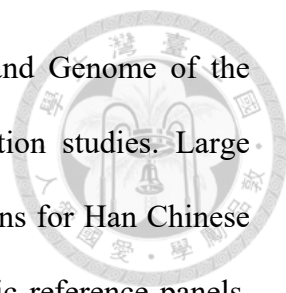
1.5 Reference panel in genotype imputation

Genotypes that are not directly assayed on GWAS arrays can be reconstructed by comparing each sample to a reference panel of sequenced genomes, in genotype imputation. In prior studies, the imputation accuracy has been found to benefit from the increase in panel size. Since the international HapMap3 project was completed in 2010 [60], more and more whole-genome sequencing (WGS) data are available to the public. The ability to impute a variant accurately is dependent both on the choice of the array and the total number of individuals genotyped in the reference panel carrying that



variant. Simulation studies have found that increasing sample size in the reference panel may improve imputation accuracy, especially for SNPs with relatively low MAFs. A large reference panel may capture many less common and rare variants, which should provide a better resolution to establish the haplotype background for observed variants [61]. At present, there exist many public reference databases. The most widely known is the 1000 Genome project, which collects the sequence data of 26 ethnic groups in the world, and divides these 26 ethnic groups into five main groups according to regions, AFR (Africa) AMR (American ethnic group) Ad Mixed American) EAS (East Asia) EUR (Europe) SAS (South Asia), a total of more than 80 million SNPs sites and 5008 haplotypes were obtained. Because of the huge amount of data and the number of ethnic groups, 1000 Genome project is now the most commonly used database of imputation reference panel. With the development of sequencing and whole genome sequence technology, there are gradually more and more large-scale reference databases, such as Haplotype Reference Consortium (HRC) and Trans-Omics for Precision Medicine (TOPMed).

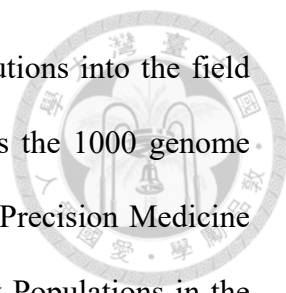
Publicly available reference genomes as mentioned above were constructed using multi-ethnic populations. However, using such multi-ethnic panels to conduct imputation, it was found that Europeans displayed the highest accuracy and Africans the lowest. Due to the fact that Asian populations possess some unique genetic characteristics, it is neither possible nor appropriate to directly adapt genetic information from studies that have been conducted for Caucasian populations). In addition, current research has also found that there are many population-specific rare SNPs [62], these SNP sites only exist in specific ethnic groups, and are often found to be related to diseases and cancers, etc. Therefore, population specific databases are



becoming more and more important. Such as the UK 10K [63] and Genome of the Netherlands [64] are emerging and used to do the local population studies. Large reference panels such as HRC have been shown to display limitations for Han Chinese populations, suggesting the necessity of building population-specific reference panels. Factors such as genetic alterations and/or mutations coupled with family history and race have been thought to play important roles in the heritability of genetically complex diseases [65]. Moreover, a higher false positive rate has been observed in imputation from global reference panels compared to imputation performed using a local panel. There is no representation of the Taiwanese population in large reference panels such as 1000 Genomes and HRC. The pan-Asian SNP genotyping database (PanSNPdb) [66], which collected SNPs and copy number variations from 1,719 samples in 71 populations including mainland China, India, Indonesia, Japan, Malaysia, the Philippines, Singapore, South Korea, Taiwan, and Thailand, also has a low Taiwanese representation. Therefore, constructing reference panels' specific for the Taiwanese population is an immediate requirement for conducting rare variant association studies in Taiwanese patients.

1.6 Imputation Server

Researchers nowadays use genotype imputation as a basic step before GWAS analysis. However, imputation requires huge computational resources and is a complex process. In addition, the usage of these imputation software requires bioinformatics expertise. Therefore, such requirements pose a bottleneck, especially for non-bioinformatics researchers. To help to solve the problem, there exist several imputation servers. One of the most popular public imputation platforms, is the

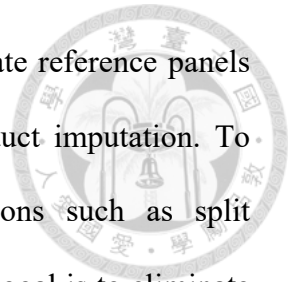


Michigan Imputation Server, which has been making great contributions into the field of population genetics research, houses reference genomes, such as the 1000 genome phase3, Haplotype Reference Consortium (HRC), Trans-Omics for Precision Medicine (TOPMed) and the Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA) [67]. The latest version of the Michigan server uses Eagle2 – Minimac4 to conduct pre-phasing and imputation, respectively, which can provide a faster speed for imputation. Another commonly used imputation server is Sanger Imputation Service, which utilizes SHAPEIT-PBWT, to let the researcher impute the data with the population specific databases UK 10K and the common reference panel 1000 Genome phase3 online. However, the primary focus of the two servers has been the European and Caucasian populations, and there are some restrictions on their use, such as the amount of data uploaded and the number of jobs used in a day. With a virtual collaboration platform, where physicians, researchers worldwide can upload their genotype data and get the required imputation done, without requiring computational resources or bioinformatics knowledge.

1.7 Specific aim

In this study, we aim to build a new public imputation and analysis platform, Multi-racial Imputation System (MI-System), to allow users access to a high speed and efficient platform to seamlessly conduct genotype imputation. The server uses the popular imputation tools SHAPEIT2-IMPUTE2 as the choice of software for pre-phasing and imputation and provides flexible quality control options for the customized imputation analysis. Moreover, we also provide a population-specific reference panel from the Taiwan biobank data (TWB), to help researchers with ethnic

specific reference panel. MI-System will further allow users to create reference panels using their customized datasets and use that as reference to conduct imputation. To ensure user convenience, the system also offers utility functions such as split chromosome and Liftover (function to convert genome builds). The goal is to eliminate the need for bioinformatics expertise, which is otherwise a pre-requisite for obtaining highly accurate imputed datasets.





Chapter 2 Materials and methods

2.1 System implementation

The architectural design of the MI-System is displayed in Figure 1. MI-System utilizes web-based Python Django to develop its back-end framework while the front-end is designed and established using HTML, CSS, and JavaScript. It divides the website functions into several python scripts that connect the website front-end to the server back-end. Django Q and Python were implemented to schedule and multi-process user uploaded jobs and output results in a first out fashion. Furthermore, to enhance computation speed of imputation and other analysis, parallel virtual machines have been designed and implemented into the server so that the program can simultaneously run on multiple regions of the chromosome for efficient processing of user uploaded data.

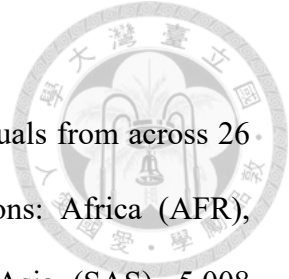
Users can upload the DNA genotype data in plink format (.bed .bim .fam) or in vcf format (.vcf.gz) using either the browser function directly, or by uploading their data in google drive and pasting the link in the allotted box in the system. For larger datasets, pasting the link to upload files is recommended, to avoid, upload breaks. The Variant Call Format (.vcf) format files will be converted to plink binary format using plink 1.9 software, for conducting all analysis. Once all analysis is accomplished GEN2VCF [68] (<https://bitbucket.org/4shin/division-of-genome-research/src/master/GEN2VCF>), a conversion software, is employed in our back-end, to convert all analyzed data from the impute2 output(.gen) format into (VCF) format, post which the -concat function of the BCFTTools [69] is used to merge each vcf file (for each imputed chunk) into one .vcf file for a chromosome. Summary plots are drawn using Python. All output files are packed into a compressed .zip folder, for the users to download with easy clicks.

2.2 MI-System Reference Panels

Mi-system includes (i) for the first time, a Taiwan Biobank (TWB) reference panel, that is created with samples of Han-Chinese origin residing in Taiwan [70], (ii) 1000 genomes Phase III reference panel, and (iii) Hapmap 3 reference panel. Users can either choose, any one panel to conduct pre-phasing and imputation or both panels and conduct imputation using the “merge reference panel” option. Furthermore, the users can customize reference panels, using the MI-System’s ‘create reference panel’ function (described later), and choose to use it as a reference, or may combine their customized panels with TWB or 1KG and use it as a combined reference, for conducting pre-phasing and imputation.

Taiwan Biobank Reference panel:

The panel is constructed using whole genome sequencing data from 997 unrelated, relatively healthy individuals randomly selected from among 20,117 Han Chinese participating in the Taiwan Biobank (TWB). The majority of the Taiwan population are of Han-Chinese ancestry and have immigrated from southeast China over the past 4 centuries, while about ~2% are of aboriginal ancestry (Austronesian). Out of the total of 997 individuals that are included in this panel, the whole genomes of 499 were sequenced using Illumina Hi-Seq 2500 and the remaining 498 were sequenced using Ion Torrent-Proton technology. A total of 26,051,907 SNVs and 3,592,314 indels with at least 30X coverage were identified where, 32.7% of total SNVs and 43.7% of total indels were found to be novel when compared to gnomAD [71]. The mean number of total and novel SNVs per individual were 1,894,528.7 (range 585,773-2,740,034), and 52,394.1 (16,037-91,629), respectively, whereas that of indels were 133,869 (31,182-306,432) and 8,746.7 (3,403-51233), respectively.



1000 Genome Reference Panel:

A reference panel created with sequenced data of 2504 individuals from across 26 global populations, that are clustered into 5 major sub-populations: Africa (AFR), America (AMR), Europe (EUR), East Asia (EAS), and South Asia (SAS). 5,008 haplotypes at over 88M variants.

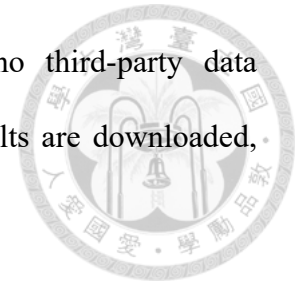
HapMap3:

The HapMap3 reference set contains genomic data from >1,000 individuals of 11 different ancestries (<https://www.sanger.ac.uk/resources/downloads/human/hapmap3.html>): African Americans (ASW), North Europeans (CEU), Chinese Americans (CHD), Gujarati (GIH), Japanese and Chinese (JPT+CHB), Luhya (LWK), Mexicans (MEX), Maasai (MKK), Toscani (TSI), and Yoruba (YRI). The panel contains approximately 1.5 million variants that are genotyped. This reference panel was included to provide users with a genotyped panel (not sequenced) as a reference. This is the only reference panel in MI-System that conforms to human genome version 18 (hg18). Therefore, HapMap3 is designated as the reference genome, and users must upload their genotype data in hg18 format to obtain the best results.

2.3 MI-System: Services

All functions and workflows of the MI-System webserver are displayed sequentially in Figure 2, and the imputation workflow shows in the Figure 3. MI-system broadly offers four different services: (i) imputation, (ii) create reference panel, (iii) split chromosome and (iv) Liftover (Figure 2a). “Imputation” and “create reference panels” are the two primary functions that allow users with an access to the inbuilt analysis pipeline for conducting imputation, and convert user uploaded datasets to

reference genome formats, respectively. The system ensures no third-party data breaches. Moreover, once the assigned jobs are done and the results are downloaded, users can choose to delete their results files in any time.



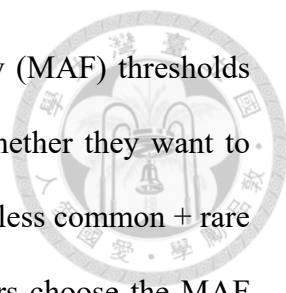
2.3.1 Service 1: Imputation

Data upload:

Users should begin by assigning a name for their project, which will be referred to, later, when the job is accomplished, for the purpose of downloading the results. To take advantage of the resources, users must upload data, one chromosome at a time, and confirm that their data adheres to human genome build 37 as all public reference panels are from hg37. The users can also determine a specific region, using the system, based on their requirement (Figure 2b). The files are required to be uploaded in either .plink binary format or .vcf files into the system.

Quality control options:

The quality control steps are conducted using Plink v1.9. It is of extreme importance to remove poor quality SNPs before proceeding for further analysis, as lesser data quality can lead to inaccurate imputations which can later produce false positives and false negatives from genome wide association analysis. For the quality control step, users can use easy clicks to choose from drop down menus, allowable thresholds, for removing poor quality SNPs (**single SNP missing rate**) and poorly genotyped individuals (**individual SNP missing rate**) from further analysis (Figure 2c). Ideally, SNPs with call-rates >90%-95% in study subjects, and individuals with <5-10% missing genotypes are retained for further analysis. A Hardy-Weinberg equilibrium (HWE) check is also required to remove SNPs that deviate from HWE. Lastly the



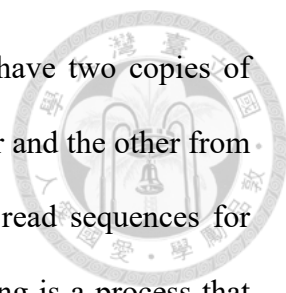
system also allows the users to choose their minor allele frequency (MAF) thresholds (0.05, 0.01, 0.001 or none). Users can filter out SNPs based on whether they want to include only common variants, common + less common, common + less common + rare variants, or all variants in their genotype data. For example, if users choose the MAF threshold as 0.05, all SNPs with $MAF < 0.05$ will be excluded from further analysis. Studies have shown that higher GWAS chip density does have a positive effect on imputation quality, however, it is also argued that rarer SNPs are usually excluded as they may have a negative effect on the accuracy, due to poor genotype calling, and lack of linkage disequilibrium (LD). Hence, the users should contemplate the pros and cons of the MAF threshold and choose it wisely with respect to their study goals.

Select study specific Reference Panel:

Once quality control steps are done, users need to select a reference panel, relevant to their study goal (Figure 2d), for accomplishing pre-phasing and imputation. The users can choose any one panel from a drop-down menu provided by Mi-System : (1) TWB reference panel, (2) 1000 genome phase III reference panel or (3) HapMap3 Reference panel, (4) a custom reference panel or (5) combine reference panel. For “custom reference panel” the users need to upload their data in the IMPUTE2 reference format which will be used solely as the reference for the rest of the pipeline. For “combine reference panel”, users can either select multiple panels provided by the system which will be merged to create a merged or improved panel, or can upload their customized data files (to create a new reference panel) and merge it with either 1000G or TWB, to be used as the combined panel.

Pre-phasing:

MI-System uses SHAPEIT2 for pre-phasing the user input genotype data in



concordance with the reference genome. Homo-sapiens (humans) have two copies of each chromosome (diploid), one of which is inherited from the father and the other from the mother. The technologies used for sequencing, generate short read sequences for individuals and post-assembly a single sequence is produced. Phasing is a process that constitutes of distinguishing the paternal and maternally obtained chromosomal strands into homologous chromosomal pairs. This requires complicated statistical computations, that uses the LD structure of the SNPs in the genotype and the reference genome to reconstruct haplotypes. This step is critical as it restores the functional consequences that was originally to be, in context to shared ancestry. From imputation's point of view, pre-phasing allows a one-time phasing where a phased genotype data can match to a selected strand of the reference panel, instead of matching two un-phased genotypes while doing imputation step, therefore speeding up the consecutive imputation steps while reducing the cost of computation. SHAPEIT2 operates in 3 steps where it first checks the alignment of the SNPs from the user uploaded data with the forward strand of the reference genome and removes the misaligned ones in the second step and finally pre-phases the user uploaded genotype data.

Imputation:

IMPUTEv2.0 software is employed in this web server to conduct imputation. Once the pre-phasing steps are done imputation commences, where each chromosome is chunked into 5 Mb regions to conduct analysis. Impute2 attaches labels to SNPs based on the panel they belong to (user uploaded genotype panel (label = 2) or reference panel (label = 0 if phased, label =1 if un-phased)). Based on the SNPs from user uploaded genotype data, corresponding matched haplotypes get selected from the reference panel and the alleles from the haplotype region are utilized to infer the missing SNP

genotypes into the user panel. The absence of observed type-2 SNPs in each of the base-pair region, fails to impute SNPs in missing sites. Therefore, each imputation chunk is flanked at either side by an internal 250kb buffer region, by IMPUTE2, to enhance the probability of type-2 SNPs in each chunk.

Combine Imputation Panels:

MI-System additionally provides the combine imputation panel function which is the unique function in IMPUTE2. The merge imputation will use the two different reference panels to improve the imputation result. There are two merge functions use in the server, “merge imputation” and “improve imputation”. The “merge imputation” use the command “-merge_ref_panels” which called “Imputation with two phased reference panels (Merge reference panels)” in IMPUTE2. In this function, the two reference panels will first imputed with each other to generate the large reference panels, then impute the sample with the big reference panels to get the high coverage result. Another function call “improve imputation”, which called “Imputation with two phased reference panels” in IMPUTE2. In this function, the sample will first impute the SNP variants that only present in the first reference panels. Then, use the second reference panels to improve the imputation accuracy of the result.

Data Output:

Once all steps are accomplished, users will be notified by the system, through an email with links to download all results (Figure 2e). The imputed data is made available to the users through compressed zip files (Figure 2f), which can be downloaded using a few simple clicks. The .zip file will contain all imputed chunks concatenated into files in .vcf format or plink format, along with summary plots, in .jpg/.png formats, for the imputed data. The figures will consist of (i) an accuracy plot, where the info-score will

be plotted across all imputed SNPs and (ii) a distribution plot, plotting the frequency against MAF for all imputed SNPs (Figure 2f). Python module, Pandas v. 1.1.1 is used to import and analyze info score files from all imputed chunks of each chromosome, and Python module Matplotlib 3.3.2 is utilized to plot and save the results to be downloaded by the user.

Validate imputation accuracy:

To verify that the imputation can correctly fill out the missing SNP sites, we conducted a simulation study by which we masked known SNP sites and imputed them and verified them. 5 variants were selected randomly from a dataset of 188 individuals on Taiwanese origin. These 5 variants were separately masked and the imputation pipeline was used to impute them back. Finally, the imputed ones were compared with the original SNPs, to determine the accuracy and correctness of imputation.

2.3.2 Service 2: Create reference panel

This service in the MI-System offers users with easy clicks to convert self-customized data into IMPUTE2–SHAPEIT2 reference panel formats. MI-System use the perl script from IMPUTE2, “vcf2impute_legend_haps.pl”, to convert the user uploaded data to the reference format. Users need to click on the, “Create Reference Panel” service from the main function menu (Figure 2a), which will redirect them to a page where they are required to assign a project name, specify the chromosome number from a drop-down menu, and upload the relevant data files either in .vcf format or Plink binary format. Once the results are ready, the user can go to result page and all reference files will be available through a .zip file for download. The .zip folder will contain 4 files pertaining to SHAPEIT2-IMPUTE2 format: all_data.sample (list of all

samples, chromosome_file.legend.gz (chromosome specific legend file), chromosome_file.hap.gz (chromosome specific haplotype file), chromosome_file_map.txt (chromosome specific map file). Users can further upload these files to customize their reference genome, or create merged reference panels for conducting pre-phasing and imputation using SHAPEIT2-IMPUTE2.

2.3.3 Service 3: Split Chromosome

This function is added to the system to help users to split their whole genome data into chromosomes. Microarray or sequence data is usually available for whole genome, and non-bioinformaticians may not be equipped enough to break down the whole genome data into chromosomes. Users are required to upload their whole genome data in .vcf or plink binary format and download the split data (chr1 – chr22, chrX) in .vcf or plink binary format. As MI-system accepts only per-chromosome data, this function would make it convenient for users to obtain suitable formats with simple clicks.

2.3.4 Service 4: Liftover

Another requirement for MI-System is that all data should be uploaded into the system in human genome version 37. Most sequencing and microarray technologies, align data to the latest release of the human genome version (HG38). However as 1000G Phase III reference genome, was aligned to human genome version HG37 (hg19), therefore for consistency users are required to ensure that all data is uploaded in the hg19 version. Thus, for the convenience of the users, the function, ‘Liftover’ (UCSC Liftover utility for switching genome assembly versions) [72] is added to the system, where users can upload their dataset in either .vcf or plink binary format and all data

will be converted from HG38 to Hg19. The Liftover is also accomplished by liftoverPlink (<https://github.com/sritchie73/liftOverPlink/>), a software that can help to convert the format between plink and the liftover file. Users can download the converted data, and consequently use other functions, offered by the system.

Computational resources

MI-system works on 4 parallel virtual machines (VM), to handle computationally intensive jobs submitted by users. Two VMs have 120 cores and 450GB memory, each, and the other two have 64 cores, and 512GB, memory each. They operate on CentOS Linux release 8.2.2004. For enhancing speed, python multi-thread processing is utilized to analyze multiple 5Mb regions of chromosomes simultaneously. For example, chromosome 1 can be divided into 50 chunks, of size 5Mb each. Each chunk uses 1 core. For a data with 2000 samples, 50 chunks when parallel processed used up 200GB and the complete imputation pipeline took 32 minutes. Imputation with merge reference panel took, 1 hour 09 minutes.



Chapter 3 Results

3.1 Web-interface:

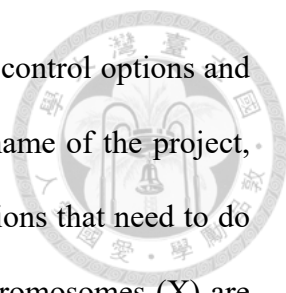
Mi-system is a python-base web system. The screenshot of the homepage is shown in Figure 4a. This page shows the basic information of the webpage and the functions offered by the web server. In order to identify the user, the webpage requires the user to enter his/her email as the user id will be used to store the result of the imputation and other functions (Figure 4b). The email will only be used for the purpose of identifying users and creating user folders. Once the user email is created, the user can select the desired function through the drop-down menu link at the top of the page.

3.2 Public reference panels

Figure 5 displays the number of SNPs that overlap between TWB panel and each of the public reference panels, 1KG and HRC. HRC panel was principally constructed of populations of European ancestry from 20 different low coverage whole genome sequencing data. The HRC contains total 39.1 million SNPs, 1KG contains total 81.7 million SNPs, and TWB contains total 39.2 million SNPs. A comparison analysis displays that approximately 9.9 million SNPs were found to overlap in all 3 panels, while, 15.9 million SNPs were common between 1KG and TWB panels, 10.2 million between TWB and HRC panels and 30.6 million SNPs between 1KG and HRC panels.

3.3 Service: Imputation

Imputation is the primary function of MI-System, and the function page is as shown in Figure 6. This page is divided into three drop down menus for users to choose



the imputation parameters from: upload data format options, quality control options and reference panel options. In the file options, the user can input the name of the project, upload the DNA genotype file, and select the chromosomes and regions that need to do the imputation. In addition to the somatic chromosomes, the sex chromosomes (X) are also available for imputation. Notice that the X chromosome will be divide into PAR1, NON-PAR and PAR2 regions to conduct the imputation. In order to avoid issues such as that the user uploaded file is too large causing the upload time to be too long, MI-System consists of a link upload function. The user is simply required to paste the google drive link and MI-System can download the files through the server back-end. In the quality control options, users can choose their own filtering thresholds, including MAF, single SNP missing rate, individual SNP missing rate, Hardy-Weinberg equilibrium check. In the reference options, users can choose the reference panel (1) TWB reference panel, (2) 1000 genome phase III reference panel, (3) HapMap3 reference panel, provided by the system or (4) upload custom reference panel for imputation.

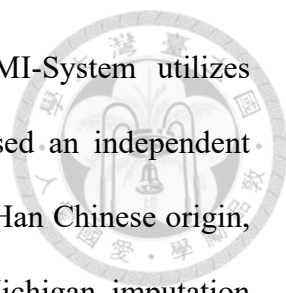
After submitting the data, the project name will show in the result page “Imputation region” (Figure 7), and the user can choose the project and move to the download page. Figure 8 shows the download page's screenshot, which records the imputation project options, progression, and result. For the user convenience, the server also provides several plots (1) Info score distribution (Figure 9), (2) Rare SNPs distribution (Figure 10), (3) Common SNPs distribution (Figure 11), to help the user quickly understand the imputation results.

3.3.1 Improve the cost of imputation time

Although IMPUTE2 is considered to be a highly accurate imputation tool, due to its algorithm and the limitation of single CPU computation, it takes more time to compute than other software. Table 2 shows the time required for imputation using IMPUTE2 on sample size data of about 100 individuals. The time necessary for imputation of chromosome 22 takes several hours, and for chromosome 1 it takes nearly a day to perform imputation. As it is not the scope of this study to modify the algorithm of the IMPUTE2 program, we use python module “threading” to allocate CPUs to different chromosomal chunks for imputation, and achieve the parallel computation of the imputation. The details of the parallel computation design are displayed in Figure 12. In the original imputation process, each chromosome is cut into several chunks of size 5 Mb, and then imputation is performed on each chunk one after another. After adding parallel computation, the complete imputation for chromosome 22 takes 30 minutes, while for chromosome 1 only takes nearly 3 hours to perform imputation, resulting in a significant improvement in the computational speed of IMPUTE2.

3.3.2 Comparison of MI-System with Michigan Imputation server.

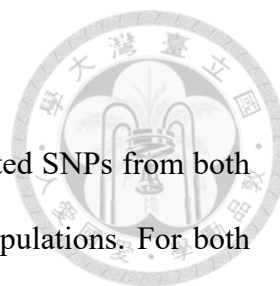
Several imputation servers exist and each uses a different combination of imputation tools thus affecting the accuracy of imputation results. For imputation, accuracy is a very important indicator that reflects the probability of calculating the correct SNP nucleotide. In order to understand whether the combination of tools used by MI-System has a better accuracy or not, we conducted a comparison analysis between MI-System and the Michigan Imputation server, to provide users with example results for gauging the reliability of MI-System. Michigan Imputation server utilizes



Eagle 2 for pre-phasing and Minimac4 for imputation, and MI-System utilizes SHAPEIT2 for pre-phasing and IMPUTE2 for imputation. We used an independent genotype dataset consisting of 94 patients (90 males, 4 females) of Han Chinese origin, living in Taiwan, to conduct a comparison analysis using the Michigan imputation server pipeline and MI-System pipeline. The data was uploaded into the Michigan imputation server and MI-System and imputation were conducted on all autosomes (chromosome 1 – chromosome 22) using 1KG as the reference panel. 1KG is particularly selected because it is the common reference panel in both the web servers, the major target population for MI-System is Asian population. HRC panel mostly works for European population, hence was not suitable to conduct comparisons. EAS population from 1KG was the major focus for all comparisons, nevertheless a separate analysis on EUR population is also done to establish the wider applicability of MI-System. A threshold of 0.01 for both SNP and individual missing-ness was used to remove poor quality SNPs and those that did not conform to HWE (threshold of 10⁻⁶), were excluded from further analysis. Finally, SNPs with MAF < 0.001 were also removed from all analyses. We conducted QC steps in MI-System using its readymade drop-down menu for QC, MAF and HWE filtering. Before feeding the clean data into the Michigan imputation server, we conducted all QC and exclusions (HWE and MAF) using a Linux system. All quality control parameters and exclusion thresholds were kept identical for both imputation pipelines to maintain data consistency.

Comparison result

Pre-phasing and imputation using the imputation pipelines from both MI-System and Michigan Imputation server were conducted. The analyzed data were downloaded and examined for data quality, skewness, and accuracy.



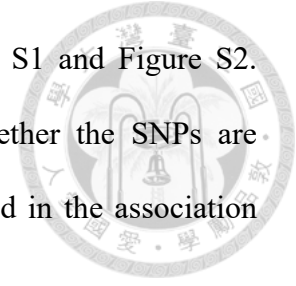
(i) **Uniform Distribution across MAFs**

Figure 13 and Figure 14 display the distribution plots for imputed SNPs from both MI-System and Michigan Imputation server, for EAS and EUR populations. For both the populations, the SNPs were mostly uniformly distributed across all MAFs. The trend of the distributions was similar for data from both the web servers. Although a quick glance revealed that the total number of SNPs with $MAF < 0.05$ were marginally higher by Minimac4, a closer look (Figure 15 and Figure 16) revealed that more number of rare and very rare SNPs ($MAF < 0.01$) were generated by IMPUTE2 and higher number of less common SNPs ($0.01 < MAF < 0.05$) were imputed by Minimac4. This distribution pattern was mostly consistent for all autosomes. We chose to display the results from chromosome 1 and chromosome 22.

(ii) **More SNPs with higher accuracy**

Figure 17 provides plots displaying accuracies of imputed SNPs using both MI-System and Michigan server, for chromosome 1 and chromosome 22. Again, all results are provided for both EAS and EUR populations, to demonstrate the multi-ethnic applicability of the imputation system. As widely understood, IMPUTE2 provides a metric, called info score as the measure of accuracy for each imputed SNP while Minimac4 provides a dosage r^2 as the measure of accuracy. Hence for the purpose of comparison, we recalculated the r^2 values (squared correlation coefficient) between imputed allele dosages and masked genotypes, by utilizing an R package “BinaryDosage” that utilized the .gen files from the impute2 output to calculate r^2 dosage values. Figure 17 shows the plotted r^2 values for both impute2 and Minimac 4, and clearly the accuracy for Impute2 was much higher than that of Minimac 4 across all MAF values. Results for chromosome1, and chromosome 22, for both EAS and EUR

shows similar trends. All other autosomes are included in Figure S1 and Figure S2. Higher accuracy for most SNPs implies that irrespective of whether the SNPs are common, rare or very rare, the probability of them getting included in the association study and consequently getting detected, becomes higher.

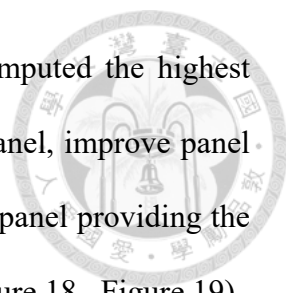


3.3.3 Merge reference panels

Rare variants are believed to contribute to the unexplained heritability of most traits and diseases, however, due to their poor representation and low frequency in populations it's been a challenge to analyze them in GWAS studies, with adequate power. This section focuses on displaying the users with examples, that merging ethnic specific panels with matched sub-population from multi ethnic panels, can enhance the representation of rare variants in population genetics studies. We conducted imputation by utilizing the “merge reference panel” and “improve reference panel” options, offered by MI-System, to combine Taiwan specific reference (TWB) with EAS of 1KG reference. We evaluated the findings in comparison to using only TWB and only EAS of 1KG based on data quality, skewness, and accuracy.

(i) Merging panels provide more imputed SNPs

We conducted imputation using 4 different reference panels, using the MI-system. “Merge” reference panel is created when 1KG EAS and TWB were merged using the “merge reference” option (Figure 2d), “improve” reference panel is created when EAS is used as the main Panel which is merged with a subset of SNPs from TWB panel that overlaps with 1KG, using the option improve reference” (Figure 2d). “EAS” is when 1KG is only used as the reference and “TWB” is when only TWB reference panel is used. Figure 16a and 16b display the distribution of SNPs across all MAFs for



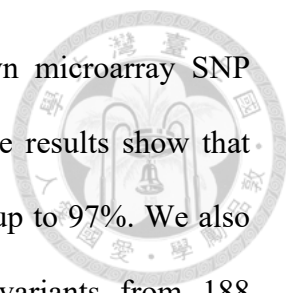
chromosome 1 and chromosome 12. “Merge” panel consistently imputed the highest number of SNPs, uniformly for all MAFs, followed by IKG EAS panel, improve panel and TWB panel. For $MAF < 0.05$, the trend was similar with Merge panel providing the highest number of SNPs in comparison to the other three panels (Figure 18 , Figure 19).

(ii) Merging panels improve the accuracy of rare SNPs.

Choosing an appropriate reference panel for the study data at hand, sample sizes of both the genotype data and the reference panels and allele frequencies, are some of the factors that have been shown to affect the quality of imputation. Figure 8 shows the comparison of info-score for all 4 panels. EAS was found to show the highest accuracy for $MAFs \geq 0.075$ (Figure 20 a-b), however for less common variants ($0.01 < MAF < 0.05$) EAS +TWB improve was shown to have higher accuracy than all others (Figure 8c-d), and for rare and very rare variants ($MAF < 0.01$), TWB reference panel showed higher accuracy than others (Figure 21 a-b). Even though for this particular example the, the enhancement of the accuracy was marginal, we believe it is due to the very low sample size of the TWB panel and genotype data. All other autosomes are included in Figure S3 and Figure S4 A customized panel created with thousands of individuals and the study data having hundreds of individuals, would allow users to conveniently merge reference panels using MI-System, which would enhance the representation of rare variants in population-based studies, thus increasing the chances of them getting detected in GWAS analyses.

3.3.4 Validate the imputation accuracy

Imputation is now often used as a processing step before analyzing GWAS, and it is important to ensure that the missing SNP variants can be estimated correctly. In the



past, many studies have compared imputation results with known microarray SNP variants data to verify the accuracy of imputation [73, 74], and the results show that imputation can estimate the missing SNP data with an accuracy of up to 97%. We also performed a similar validation by randomly masking 5 SNP variants from 188 individuals from Taiwan and then conducted imputation with 1KG EAS group. The imputation results were compared with the original data to calculate the percentage of the correctly imputed genotypes. The results show that the mean accuracy rate of the imputed data were up to 99% (Table 1), therefore it can be verified that the imputation process is able to accurately fill out the missing SNP variants and improve the further GWAS analysis.

3.4 Service: Split chromosome

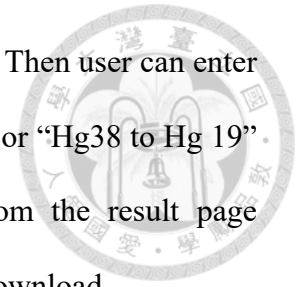
To optimize the operation of the server, MI-System limits users to imputation of only one chromosome at a time. Therefore, we additionally provide split chromosome function that allows users to upload their whole genome SNPs data to the server and split it into single chromosome data from chr1 to chr22 via the PLINK software. The screenshot of this utility is displayed in Figure 22. Users can upload files via browser or google drive link, and customize the name of the output file. After the users submit their data, the file will be sent to the server back-end to start the program. The results can be downloaded by going to the download page “Split chromosome region” selecting either plink format or vcf format for download.

3.5 Service: Liftover

Since the Human Genome Project began in 1990, many versions of the human genome assemblies have been released, the most commonly used version is GRCH37 (hg19), which was released in 2009, and GRCH38 (hg38), which was released in 2013. Currently, major genomic databases such as NCBI, UCSC, Ensembl, 1000 Genomes Project, and gnomAD can use these two versions for searching and data usage. However, there are still some extended tools that have not been fully updated. Therefore, for now, using the GRCH37(hg19) can better support the research application. Moreover, the genome coordinates and genome annotation are different between the different versions of the human genome assemblies. The genome data between two different versions cannot be directly used for further analysis.

To solve these issues, UCSC has published a conversion tool, liftOver (<https://genome.ucsc.edu/cgi-bin/hgLiftOver>), that allows users to convert genome coordinates and genome annotation files from one human genome version to another. Several of the reference databases that exist today still use data in GRCH37 (hg19) version, and the associated analysis software is also compatible with HG37. However, depending on the sequencing protocol used, the study SNP data may pertain to human genome version 38 GRCH38 (hg38). Also, if users use the online liftover conversion tool from UCSC browser, they must create a specific format of input data, as the online version of liftover does not support Plink files. MI-System has incorporated liftOverPlink to help convert the genome coordinates version using plink format. LiftOverPlink first converts the input plink format data to .bed file, which is then converted using LiftOver, finally producing the output in the plink format. Figure 23 displays the screenshot of the liftover page. The page can let the user upload the data in

plink format or vcf format through the browser or google drive link. Then user can enter the output name and the conversion options, either “Hg19 to Hg38” or “Hg38 to Hg 19” based on their requirement. The results can be downloaded from the result page “liftover region” by selecting either plink format or vcf format for download.

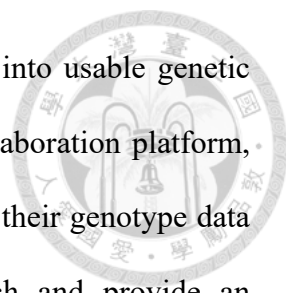


3.6 Create reference

Currently, many reference datasets exist, such as 1000 genome and HRC. These large reference databases can have high imputation accuracy when utilized for imputation analysis. Recently, many studies have shown that conducting imputation analysis using ethnically matched reference panels can provide better performance. Some rare SNPs exist only in their population-specific databases, and these SNPs have been found to be associated with some diseases and traits. However, not every population has its own population specific reference database. MI-System offers users with a readily available function that can convert user uploaded data to be converted to reference format (.legned, .hap) for conducting imputation. MI-System allows users to upload SNPs data in vcf or plink formats (.vcf.gz, .bed, .bim, .fam), and help to convert them into reference panel format by the IMPUTE2 script “vcf2impute_legend_haps.pl”. Figure 24 shows the screenshot of the Create reference page. Here the user can upload the data, and select the chromosome number that want to convert to the reference format. The result can be confirmed by redirecting to the download page from the result page “Reference region” and selecting plink format or vcf format for download.

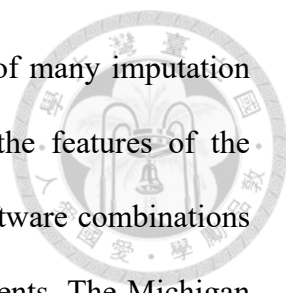
Chapter 4 Discussion

Physicians or biologists, who possess large-scale patient data, lack the expertise



and the computational resources required for translating their data into usable genetic information for further research. Providing them with a virtual collaboration platform, where physicians and researchers worldwide can seamlessly upload their genotype data and get the required imputation done, would speed up research and provide an unfathomable genetic knowledge base for genetic carriers of complex diseases, therefore greatly enhancing patient driven research. We release an easy to operate webserver, MI-System, where users can choose QC parameters for conducting data cleaning, and flexibly conduct imputation using either public imputation panels or using self-customized/self-merge multiple panels, using few easy clicks. Moreover, once the user-designed jobs are accomplished, the system provides visual representations of the data quality through accuracy distribution plots and minor allele frequency (MAF) distribution plots. Users can instantly judge the quality of their imputed data. This further saves the user, of complex post-processing steps, such as merging imputation chunks, or writing additional codes for data plotting and checking overall imputation quality.

Genotype imputations are becoming increasingly popular as it helps to enhance the representation of rare variations in population-based studies. This is of utmost importance as rare variants originally explain risks in very few people in population-based studies. Therefore, it is challenging to explain their effects in the greater population. Furthermore, it is hard to assay them as tagged single-nucleotide polymorphisms (SNPs) in genome-wide arrays, unless the sample size is very large. Also, assaying rare variants in microarrays is expensive. Due to such bottlenecks, rare variants never make it to association studies and get excluded at quality control steps. Even if they get included, they hardly show association, because of very low effect size.



The demand to conduct imputation has triggered the development of many imputation panels and servers. Table 3 gives a detailed comparison of all the features of the existing panels. Each online imputation server uses the different software combinations and reference panels for the imputation according to their requirements. The Michigan imputation server uses Eagle2-minimac4 for pre-phasing and imputation, Sanger imputation service uses SHAPEIT-PBWT for pre-phasing and imputation, respectively. Although IMPUTE2 is considered to have high imputation accuracy, its single CPU computing limitation causes long computation time and is therefore seldom used in online servers. About the reference panels that servers use, except for the 1000 Genome Phase3 panel, the other panels were mainly for European and Caucasian populations. Therefore, MI-System provides the high accuracy software combination SHAPEIT-IMPUTE2 to do the pre-phasing and imputation. Further, MI-System has added a new Taiwan biobank reference panel and has provided options for customizing user specified reference panels. Compared with other online servers, MI-System takes more time for imputation, but it can get better imputation accuracy and has the special merge reference panel function. Nowadays, there are new versions of SHAPEIT-IMPUTE2 used by MI-System, SHAPEIT4-IMPUTE5 [75, 76], which update the algorithms and functions used. Not only does it remove the limitations of single CPU computing, but it also reduces memory consumption while maintaining high accuracy. Therefore, we have also added an imputation page for the SHAPEIT4-IMPUTE5 version in the server, allowing users to choose. With the addition of new supported reference panels and tools in the future, we believe that the MI-System can help more researchers to conduct imputation and further SNP studies.

Chapter 5 Conclusions

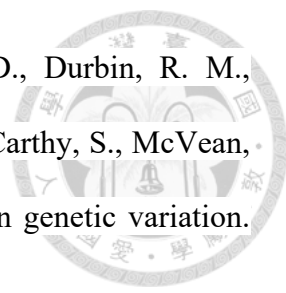
MI-System is a new online web-based server that can help users conduct imputation. It provides popular and high-accuracy imputation tools SHAPEIT-IMPUTE2 for pre-phasing and imputation, respectively. Besides the publicly available 1000 genome Phase III reference panel, a new Taiwan biobank reference panel is added for the first time in this server. The imputation function can let the user flexibility select the quality control parameters and reference panel options. Users can also use the combined reference panels function to improve the imputation accuracy. Moreover, MI-System provides several useful functions, (1) Split chromosome, (2) LiftOver, (3) Create reference, to help users to do the analysis and improve the user experience. By using MI-System, we hope that users can get imputation results quickly and help with further analysis.

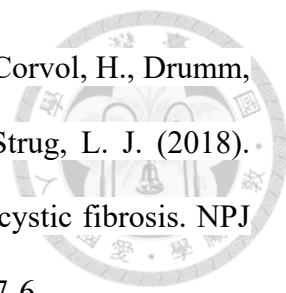


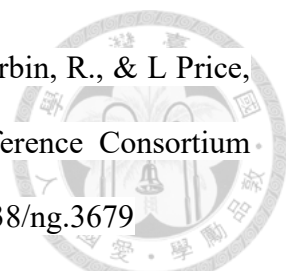


Chapter 6 References


1. Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorf, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., ... Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265), 747–753. <https://doi.org/10.1038/nature08494>
2. LaFramboise T. (2009). Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic acids research*, 37(13), 4181–4193. <https://doi.org/10.1093/nar/gkp552>
3. Pierre, A. S., & Genin, E. (2014). How important are rare variants in common disease? *Briefings in Functional Genomics*, 13(5), 353-361. [doi:10.1093/bfpg/elu025](https://doi.org/10.1093/bfpg/elu025)
4. Li, Y., Willer, C., Sanna, S., & Abecasis, G. (2009). Genotype imputation. *Annual review of genomics and human genetics*, 10, 387–406. <https://doi.org/10.1146/annurev.genom.9.081307.164242>
5. Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7), 499-511. [doi:10.1038/nrg2796](https://doi.org/10.1038/nrg2796)
6. Higgins, J. P., White, I. R., & Wood, A. M. (2008). Imputation methods for missing outcome data in meta-analysis of clinical trials. *Clinical trials (London, England)*, 5(3), 225–239. <https://doi.org/10.1177/1740774508091600>
7. Davidovich, O., Halperin, E., Kimmel, G., & Shamir, R. (2009). Increasing the power of association studies by imputation-based sparse tag SNP selection. *Communications in Information and Systems*, 9(3), 269-282.

- 
8. 1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, 526(7571), 68–74. <https://doi.org/10.1038/nature15393>
 9. International HapMap Consortium (2003). The International HapMap Project. *Nature*, 426(6968), 789–796. <https://doi.org/10.1038/nature02168>
 10. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., Kang, H. M., Fuchsberger, C., Danecek, P., Sharp, K., Luo, Y., Sidore, C., Kwong, A., Timpson, N., Koskinen, S., Vrieze, S., Scott, L. J., Zhang, H., Mahajan, A., Veldink, J., ... Haplotype Reference Consortium (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics*, 48(10), 1279–1283. <https://doi.org/10.1038/ng.3643>
 11. NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium (2021). Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, 590(7845), 290-299. <https://doi.org/10.1038/s41586-021-03205-y>
 12. Deelen, P., Menelaou, A., van Leeuwen, E. M., Kanterakis, A., van Dijk, F., Medina-Gomez, C., Francioli, L. C., Hottenga, J. J., Karssen, L. C., Estrada, K., Kreiner-Møller, E., Rivadeneira, F., van Setten, J., Gutierrez-Achury, J., Westra, H. J., Franke, L., van Enkevort, D., Dijkstra, M., Byelas, H., van Duijn, C. M., ... Swertz, M. A. (2014). Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'. *European journal of human genetics: EJHG*, 22(11), 1321–1326. <https://doi.org/10.1038/ejhg.2014.19>
 13. Panjwani, N., Xiao, B., Xu, L., Gong, J., Keenan, K., Lin, F., He, G., Baskurt, Z.,

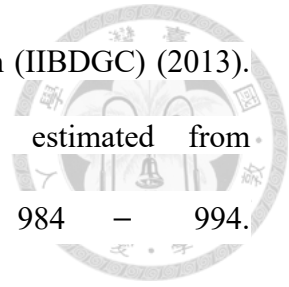
- 
- Kim, S., Zhang, L., Esmacili, M., Blackman, S., Scherer, S. W., Corvol, H., Drumm, M., Knowles, M., Cutting, G., Rommens, J. M., Sun, L., & Strug, L. J. (2018). Improving imputation in disease-relevant regions: lessons from cystic fibrosis. *NPJ genomic medicine*, 3, 8. <https://doi.org/10.1038/s41525-018-0047-6>
14. Panjwani, N., Xiao, B., Xu, L., Gong, J., Keenan, K., Lin, F., He, G., Baskurt, Z., Kim, S., Zhang, L., Esmacili, M., Blackman, S., Scherer, S. W., Corvol, H., Drumm, M., Knowles, M., Cutting, G., Rommens, J. M., Sun, L., & Strug, L. J. (2018). Improving imputation in disease-relevant regions: lessons from cystic fibrosis. *NPJ genomic medicine*, 3, 8. <https://doi.org/10.1038/s41525-018-0047-6>
15. Dang, H., Gallins, P. J., Pace, R. G., Guo, X. L., Stonebraker, J. R., Corvol, H., Cutting, G. R., Drumm, M. L., Strug, L. J., Knowles, M. R., & O'Neal, W. K. (2016). Novel variation at chr11p13 associated with cystic fibrosis lung disease severity. *Human genome variation*, 3, 16020. <https://doi.org/10.1038/hgv.2016.20>
16. Verma, S. S., Andrade, M. D., Tromp, G., Kuivaniemi, H., Pugh, E., Namjou-Khales, B., . . . Ritchie, M. D. (2014). Imputation and quality control steps for combining multiple genome-wide datasets. *Frontiers in Genetics*, 5. doi:10.3389/fgene.2014.00370
17. Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., Schlessinger, D., Stambolian, D., Loh, P. R., Iacono, W. G., Swaroop, A., Scott, L. J., Cucca, F., Kronenberg, F., Boehnke, M., Abecasis, G. R., . . . Fuchsberger, C. (2016). Next-generation genotype imputation service and methods. *Nature genetics*, 48(10), 1284–1287. <https://doi.org/10.1038/ng.3656>
18. Loh, P. R., Danecek, P., Palamara, P. F., Fuchsberger, C., A Reshef, Y., K Finucane,

- 
- H., Schoenherr, S., Forer, L., McCarthy, S., Abecasis, G. R., Durbin, R., & L Price, A. (2016). Reference-based phasing using the Haplotype Reference Consortium panel. *Nature genetics*, 48(11), 1443–1448. <https://doi.org/10.1038/ng.3679>
19. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., & Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature genetics*, 44(8), 955–959. <https://doi.org/10.1038/ng.2354>
20. O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., Traglia, M., Huang, J., Huffman, J. E., Rudan, I., McQuillan, R., Fraser, R. M., Campbell, H., Polasek, O., Asiki, G., Ekoru, K., Hayward, C., Wright, A. F., Vitart, V., Navarro, P., ... Marchini, J. (2014). A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS genetics*, 10(4), e1004234. <https://doi.org/10.1371/journal.pgen.1004234>
21. Howie, B. N., Donnelly, P., & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics*, 5(6), e1000529. <https://doi.org/10.1371/journal.pgen.1000529>
22. Karki, R., Pandya, D., Elston, R. C., & Ferlini, C. (2015). Defining "mutation" and "polymorphism" in the era of personal genomics. *BMC medical genomics*, 8, 37. <https://doi.org/10.1186/s12920-015-0115-z>
23. Shastry B. S. (2009). SNPs: impact on gene function and phenotype. *Methods in molecular biology* (Clifton, N.J.), 578, 3–22. https://doi.org/10.1007/978-1-60327-411-1_1
24. Meyer, O. S., Lunn, M., Garcia, S. L., Kjærbye, A. B., Morling, N., Børsting, C., & Andersen, J. D. (2020). Association between brown eye colour in rs12913832:GG individuals and SNPs in TYR, TYRP1, and SLC24A4. *PloS one*, 15(9), e0239131.

<https://doi.org/10.1371/journal.pone.0239131>

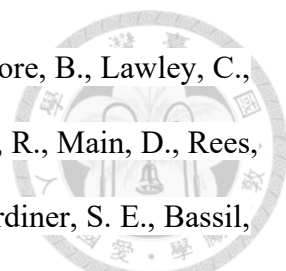
- 
25. Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., & Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature genetics*, 42(7), 565–569. <https://doi.org/10.1038/ng.608>
26. Pang, G. S., Wang, J., Wang, Z., & Lee, C. G. (2009). Predicting potentially functional SNPs in drug-response genes. *Pharmacogenomics*, 10(4), 639–653. <https://doi.org/10.2217/pgs.09.12>
27. Schlauch, K. A., Kulick, D., Subramanian, K., De Meirleir, K. L., Palotás, A., & Lombardi, V. C. (2019). Single-nucleotide polymorphisms in a cohort of significantly obese women without cardiometabolic diseases. *International journal of obesity (2005)*, 43(2), 253–262. <https://doi.org/10.1038/s41366-018-0181-3>
28. Al-Daghri, N. M., Al-Attas, O. S., Krishnaswamy, S., Mohammed, A. K., Alenad, A. M., Chrousos, G. P., & Alokail, M. S. (2015). Association of Type 2 Diabetes Mellitus related SNP genotypes with altered serum adipokine levels and metabolic syndrome phenotypes. *International journal of clinical and experimental medicine*, 8(3), 4464–4471.
29. Gregersen, P. K., & Olsson, L. M. (2009). Recent advances in the genetics of autoimmune disease. *Annual review of immunology*, 27, 363–391. <https://doi.org/10.1146/annurev.immunol.021908.132653>
30. Cross-Disorder Group of the Psychiatric Genomics Consortium, Lee, S. H., Ripke, S., Neale, B. M., Faraone, S. V., Purcell, S. M., Perlis, R. H., Mowry, B. J., Thapar, A., Goddard, M. E., Witte, J. S., Absher, D., Agartz, I., Akil, H., Amin, F., Andreassen, O. A., Anjorin, A., Anney, R., Anttila, V., Arking, D. E., ...

International Inflammatory Bowel Disease Genetics Consortium (IIBDGC) (2013). Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nature genetics*, 45(9), 984 – 994.

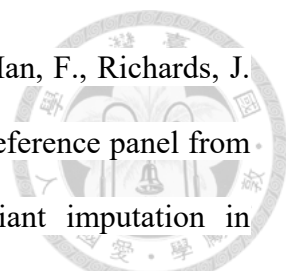


<https://doi.org/10.1038/ng.2711>

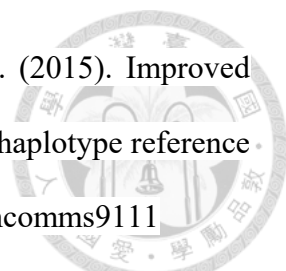
31. Deng, N., Zhou, H., Fan, H., & Yuan, Y. (2017). Single nucleotide polymorphisms and cancer susceptibility. *Oncotarget*, 8(66), 110635–110649. <https://doi.org/10.18632/oncotarget.22372>
32. Bisgin, A., Sonmezler, O., Boga, I., & Yilmaz, M. (2021). The impact of rare and low-frequency genetic variants in common variable immunodeficiency (CVID). *Scientific reports*, 11(1), 8308. <https://doi.org/10.1038/s41598-021-87898-1>
33. Kent J. W., Jr (2011). Rare variants, common markers: synthetic association and beyond. *Genetic epidemiology*, 35 Suppl 1(Suppl 1), S80–S84. <https://doi.org/10.1002/gepi.20655>
34. Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *American journal of human genetics*, 90(1), 7–24. <https://doi.org/10.1016/j.ajhg.2011.11.029>
35. Korte, A., & Farlow, A. (2013). The advantages and limitations of trait analysis with GWAS: a review. *Plant methods*, 9, 29. <https://doi.org/10.1186/1746-4811-9-29>
36. Cai, M., Dai, S., Chen, W., Xia, C., Lu, L., Dai, S., Qi, J., Wang, M., Wang, M., Zhou, L., Lei, F., Zuo, T., Zeng, H., & Zhao, X. (2017). Environmental factors, seven GWAS-identified susceptibility loci, and risk of gastric cancer and its precursors in a Chinese population. *Cancer medicine*, 6(3), 708–720. <https://doi.org/10.1002/cam4.1038>


- 
37. Chagné, D., Crowhurst, R. N., Troggio, M., Davey, M. W., Gilmore, B., Lawley, C., Vanderzande, S., Hellens, R. P., Kumar, S., Cestaro, A., Velasco, R., Main, D., Rees, J. D., Iezzoni, A., Mockler, T., Wilhelm, L., Van de Weg, E., Gardiner, S. E., Bassil, N., & Peace, C. (2012). Genome-wide SNP detection, validation, and development of an 8K SNP array for apple. *PloS one*, 7(2), e31745. <https://doi.org/10.1371/journal.pone.0031745>
38. Wang, H., Xu, X., Vieira, F. G., Xiao, Y., Li, Z., Wang, J., Nielsen, R., & Chu, C. (2016). The Power of Inbreeding: NGS-Based GWAS of Rice Reveals Convergent Evolution during Rice Domestication. *Molecular plant*, 9(7), 975–985. <https://doi.org/10.1016/j.molp.2016.04.018>
39. Louhelainen J. (2016). SNP Arrays. *Microarrays (Basel, Switzerland)*, 5(4), 27. <https://doi.org/10.3390/microarrays5040027>
40. Kumar, S., Banks, T. W., & Cloutier, S. (2012). SNP Discovery through Next-Generation Sequencing and Its Applications. *International journal of plant genomics*, 2012, 831460. <https://doi.org/10.1155/2012/831460>
41. Peng, Z., Zhao, Z., Clevenger, J. P., Chu, Y., Paudel, D., Ozias-Akins, P., & Wang, J. (2020). Comparison of SNP Calling Pipelines and NGS Platforms to Predict the Genomic Regions Harboring Candidate Genes for Nodulation in Cultivated Peanut. *Frontiers in genetics*, 11, 222. <https://doi.org/10.3389/fgene.2020.00222>
42. Roh, S. W., Abell, G. C., Kim, K. H., Nam, Y. D., & Bae, J. W. (2010). Comparing microarrays and next-generation sequencing technologies for microbial ecology research. *Trends in biotechnology*, 28(6), 291–299. <https://doi.org/10.1016/j.tibtech.2010.03.001>
43. Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide

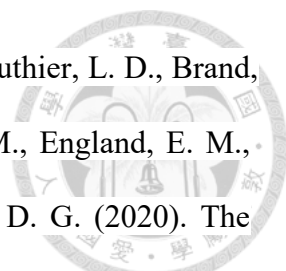
- association studies. *Nature reviews. Genetics*, 11(7), 499–511.
<https://doi.org/10.1038/nrg2796>
44. Quick, C., Anugu, P., Musani, S., Weiss, S. T., Burchard, E. G., White, M. J., Keys, K. L., Cucca, F., Sidore, C., Boehnke, M., & Fuchsberger, C. (2020). Sequencing and imputation in GWAS: Cost-effective strategies to increase power and genomic coverage across diverse populations. *Genetic epidemiology*, 44(6), 537–549.
<https://doi.org/10.1002/gepi.22326>
45. Huo, Y., Li, S., Liu, J., Li, X., & Luo, X. J. (2019). Functional genomics reveal gene regulatory mechanisms underlying schizophrenia risk. *Nature communications*, 10(1), 670. <https://doi.org/10.1038/s41467-019-08666-4>
46. Naj A. C. (2019). Genotype Imputation in Genome-Wide Association Studies. *Current protocols in human genetics*, 102(1), e84. <https://doi.org/10.1002/cphg.84>
47. Chagnon, M., O'Loughlin, J., Engert, J. C., Karp, I., & Sylvestre, M. P. (2018). Missing single nucleotide polymorphisms in Genetic Risk Scores: A simulation study. *PloS one*, 13(7), e0200630. <https://doi.org/10.1371/journal.pone.0200630>
48. Neale B. M. (2010). Introduction to linkage disequilibrium, the HapMap, and imputation. *Cold Spring Harbor protocols*, 2010(3), pdb.top74.
<https://doi.org/10.1101/pdb.top74>
49. Stephens, M., & Scheet, P. (2005). Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *American journal of human genetics*, 76(3), 449–462. <https://doi.org/10.1086/428594>
50. Chattopadhyay, A., & Lu, T. (2020). Overcoming the challenges of imputation of rare variants in a Taiwanese cohort. *Translational Cancer Research*, 9(7), 4065-4069.
doi:10.21037/tcr-20-2395

- 
51. Chou, W. C., Zheng, H. F., Cheng, C. H., Yan, H., Wang, L., Han, F., Richards, J. B., Karasik, D., Kiel, D. P., & Hsu, Y. H. (2016). A combined reference panel from the 1000 Genomes and UK10K projects improved rare variant imputation in European and Chinese samples. *Scientific reports*, 6, 39313. <https://doi.org/10.1038/srep39313>
52. Sariya, S., Lee, J. H., Mayeux, R., Vardarajan, B. N., Reyes-Dumeyer, D., Manly, J. J., Brickman, A. M., Lantigua, R., Medrano, M., Jimenez-Velazquez, I. Z., & Tosto, G. (2019). Rare Variants Imputation in Admixed Populations: Comparison Across Reference Panels and Bioinformatics Tools. *Frontiers in genetics*, 10, 239. <https://doi.org/10.3389/fgene.2019.00239>
53. Verma, S. S., de Andrade, M., Tromp, G., Kuivaniemi, H., Pugh, E., Namjou-Khales, B., Mukherjee, S., Jarvik, G. P., Kottyan, L. C., Burt, A., Bradford, Y., Armstrong, G. D., Derr, K., Crawford, D. C., Haines, J. L., Li, R., Crosslin, D., & Ritchie, M. D. (2014). Imputation and quality control steps for combining multiple genome-wide datasets. *Frontiers in genetics*, 5, 370. <https://doi.org/10.3389/fgene.2014.00370>
54. Southam, L., Panoutsopoulou, K., Rayner, N. W., Chapman, K., Durrant, C., Ferreira, T., Arden, N., Carr, A., Deloukas, P., Doherty, M., Loughlin, J., McCaskie, A., Ollier, W. E., Ralston, S., Spector, T. D., Valdes, A. M., Wallis, G. A., Wilkinson, J. M., arcOGEN consortium, Marchini, J., ... Zeggini, E. (2011). The effect of genome-wide association scan quality control on imputation outcome for common variants. *European journal of human genetics : EJHG*, 19(5), 610–614. <https://doi.org/10.1038/ejhg.2010.242>
55. Cai, C., Zhu, G., Zhang, T., & Guo, W. (2017). High-density 80 K SNP array is a

- powerful tool for genotyping *G. hirsutum* accessions and genome analysis. *BMC genomics*, 18(1), 654. <https://doi.org/10.1186/s12864-017-4062-2>
56. Wagner M. J. (2013). Rare-variant genome-wide association studies: a new frontier in genetic analysis of complex traits. *Pharmacogenomics*, 14(4), 413–424. <https://doi.org/10.2217/pgs.13.36>
57. Chen, B., Cole, J. W., & Grond-Ginsbach, C. (2017). Departure from Hardy Weinberg Equilibrium and Genotyping Error. *Frontiers in genetics*, 8, 167. <https://doi.org/10.3389/fgene.2017.00167>
58. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, 81(3), 559–575. <https://doi.org/10.1086/519795>
59. Nettelblad C. (2013). Breakdown of methods for phasing and imputation in the presence of double genotype sharing. *PloS one*, 8(3), e60354. <https://doi.org/10.1371/journal.pone.0060354>
60. International HapMap 3 Consortium, Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., Peltonen, L., Dermitzakis, E., Schaffner, S. F., Yu, F., Peltonen, L., Dermitzakis, E., Bonnen, P. E., Altshuler, D. M., Gibbs, R. A., de Bakker, P. I., Deloukas, P., Gabriel, S. B., Gwilliam, R., Hunt, S., ... McEwen, J. E. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311), 52–58. <https://doi.org/10.1038/nature09298>
61. Huang, J., Howie, B., McCarthy, S., Memari, Y., Walter, K., Min, J. L., Danecek, P., Malerba, G., Trabetti, E., Zheng, H. F., UK10K Consortium, Gambaro, G., Richards,

- 
- J. B., Durbin, R., Timpson, N. J., Marchini, J., & Soranzo, N. (2015). Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nature communications*, 6, 8111. <https://doi.org/10.1038/ncomms9111>
62. Choudhury, A., Hazelhurst, S., Meintjes, A., Achinike-Oduaran, O., Aron, S., Gamielien, J., Jalali Sefid Dashti, M., Mulder, N., Tiffin, N., & Ramsay, M. (2014). Population-specific common SNPs reflect demographic histories and highlight regions of genomic plasticity with functional relevance. *BMC genomics*, 15(1), 437. <https://doi.org/10.1186/1471-2164-15-437>
63. UK10K Consortium, Walter, K., Min, J. L., Huang, J., Crooks, L., Memari, Y., McCarthy, S., Perry, J. R., Xu, C., Futema, M., Lawson, D., Iotchkova, V., Schiffels, S., Hendricks, A. E., Danecek, P., Li, R., Floyd, J., Wain, L. V., Barroso, I., Humphries, S. E., ... Soranzo, N. (2015). The UK10K project identifies rare variants in health and disease. *Nature*, 526(7571), 82–90. <https://doi.org/10.1038/nature14962>
64. Boomsma, D. I., Wijmenga, C., Slagboom, E. P., Swertz, M. A., Karssen, L. C., Abdellaoui, A., Ye, K., Guryev, V., Vermaat, M., van Dijk, F., Francioli, L. C., Hottenga, J. J., Laros, J. F., Li, Q., Li, Y., Cao, H., Chen, R., Du, Y., Li, N., Cao, S., ... van Duijn, C. M. (2014). The Genome of the Netherlands: design, and project goals. *European journal of human genetics: EJHG*, 22(2), 221–227. <https://doi.org/10.1038/ejhg.2013.118>
65. Klein, C., & Westenberger, A. (2012). Genetics of Parkinson's disease. *Cold Spring Harbor perspectives in medicine*, 2(1), a008888. <https://doi.org/10.1101/cshperspect.a008888>
66. Ngamphiw, C., Assawamakin, A., Xu, S., Shaw, P. J., Yang, J. O., Ghang, H., Bhak,

- 
- J., Liu, E., Tongshima, S., & HUGO Pan-Asian SNP Consortium (2011). PanSNPdb: the Pan-Asian SNP genotyping database. *PloS one*, 6(6), e21451. <https://doi.org/10.1371/journal.pone.0021451>
67. Mathias, R. A., Taub, M. A., Gignoux, C. R., Fu, W., Musharoff, S., O'Connor, T. D., Vergara, C., Torgerson, D. G., Pino-Yanes, M., Shringarpure, S. S., Huang, L., Rafaels, N., Boorgula, M. P., Johnston, H. R., Ortega, V. E., Levin, A. M., Song, W., Torres, R., Padhukasahasram, B., Eng, C., ... Barnes, K. C. (2016). A continuum of admixture in the Western Hemisphere revealed by the African Diaspora genome. *Nature communications*, 7, 12522. <https://doi.org/10.1038/ncomms12522>
68. Shin, D. M., Hwang, M. Y., Kim, B. J., Ryu, K. H., & Kim, Y. J. (2020). GEN2VCF: a converter for human genome imputation output format to VCF format. *Genes & genomics*, 42(10), 1163–1168. <https://doi.org/10.1007/s13258-020-00982-0>
69. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* (Oxford, England), 25(16), 2078– 2079. <https://doi.org/10.1093/bioinformatics/btp352>
70. Wei, C. Y., Yang, J. H., Yeh, E. C., Tsai, M. F., Kao, H. J., Lo, C. Z., Chang, L. P., Lin, W. J., Hsieh, F. J., Belsare, S., Bhaskar, A., Su, M. W., Lee, T. C., Lin, Y. L., Liu, F. T., Shen, C. Y., Li, L. H., Chen, C. H., Wall, J. D., Wu, J. Y., ... Kwok, P. Y. (2021). Genetic profiles of 103,106 individuals in the Taiwan Biobank provide insights into the health and history of Han Chinese. *NPJ genomic medicine*, 6(1), 10. <https://doi.org/10.1038/s41525-021-00178-9>
71. Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q.,

- 
- Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England, E. M., Seaby, E. G., Kosmicki, J. A., Walters, R. K., ... MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809), 434–443. <https://doi.org/10.1038/s41586-020-2308-7>
72. Kuhn, R. M., Haussler, D., & Kent, W. J. (2013). The UCSC genome browser and associated tools. *Briefings in bioinformatics*, 14(2), 144–161. <https://doi.org/10.1093/bib/bbs038>
73. Juang, J. J., Lu, T. P., Su, M. W., Lin, C. W., Yang, J. H., Chu, H. W., Chen, C. H., Hsiao, Y. W., Lee, C. Y., Chiang, L. M., Yu, Q. Y., Hsiao, C. K., Chen, C. J., Wu, P. E., Pai, C. H., Chuang, E. Y., & Shen, C. Y. (2020). Rare variants discovery by extensive whole-genome sequencing of the Han Chinese population in Taiwan: Applications to cardiovascular medicine. *Journal of advanced research*, 30, 147–158. <https://doi.org/10.1016/j.jare.2020.12.003>
74. Ma, P., Brøndum, R. F., Zhang, Q., Lund, M. S., & Su, G. (2013). Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish Red Cattle. *Journal of dairy science*, 96(7), 4666–4677. <https://doi.org/10.3168/jds.2012-6316>
75. Delaneau, O., Zagury, J. F., Robinson, M. R., Marchini, J. L., & Dermitzakis, E. T. (2019). Accurate, scalable and integrative haplotype estimation. *Nature communications*, 10(1), 5436. <https://doi.org/10.1038/s41467-019-13225-y>
76. Rubinacci, S., Delaneau, O., & Marchini, J. (2020). Genotype imputation using the Positional Burrows Wheeler Transform. *PLoS genetics*, 16(11), e1009049. <https://doi.org/10.1371/journal.pgen.1009049>

Appendix

Figure:

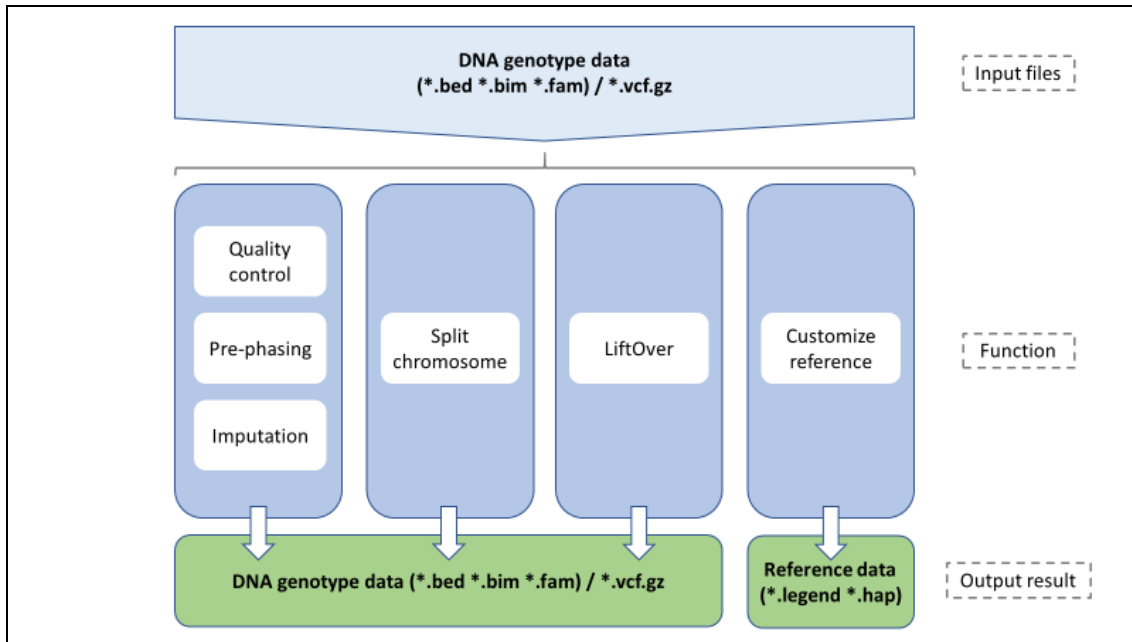


Figure1. The architectural design of the MI-System

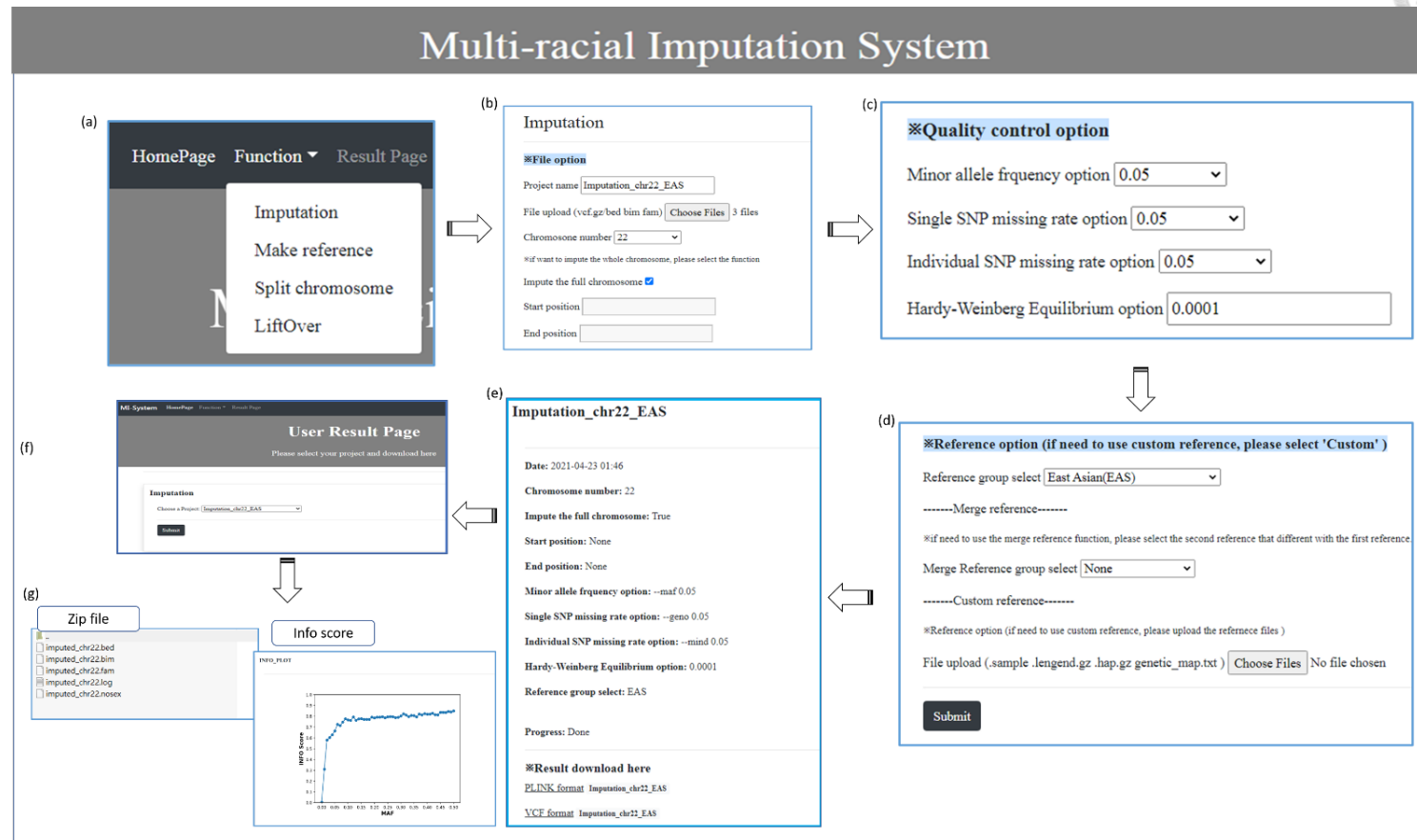


Figure 2 The overview workflow of the MI-System

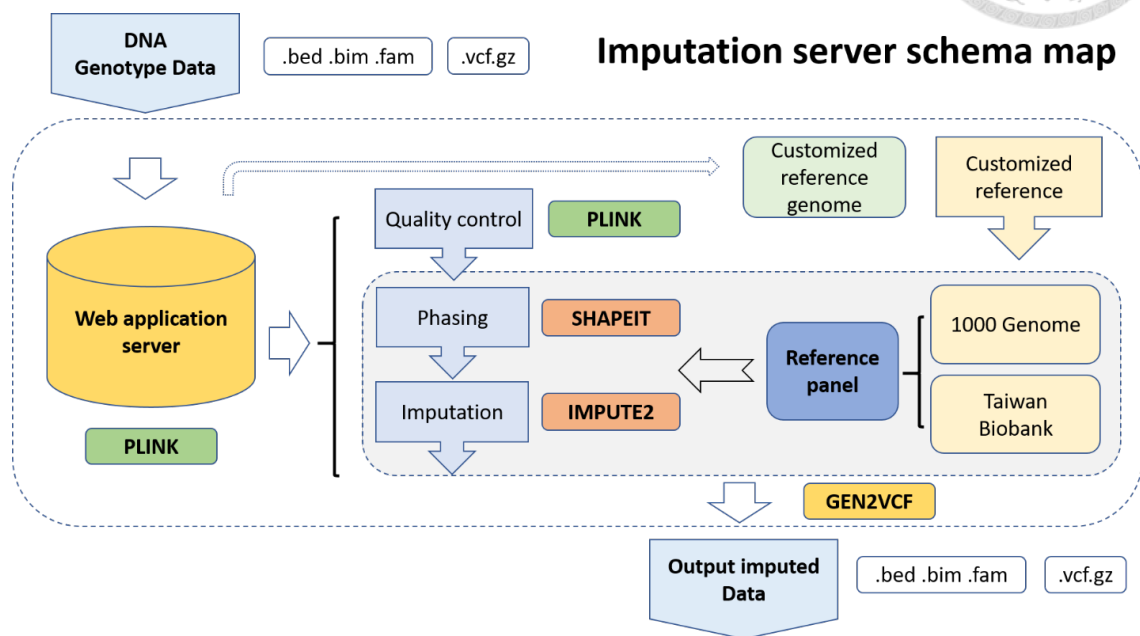


Figure 3 The workflow of the imputation function in MI-System



Enter your email as user id

Email*

Submit

Figure 4 The homepage and the email entering window of the MI-System

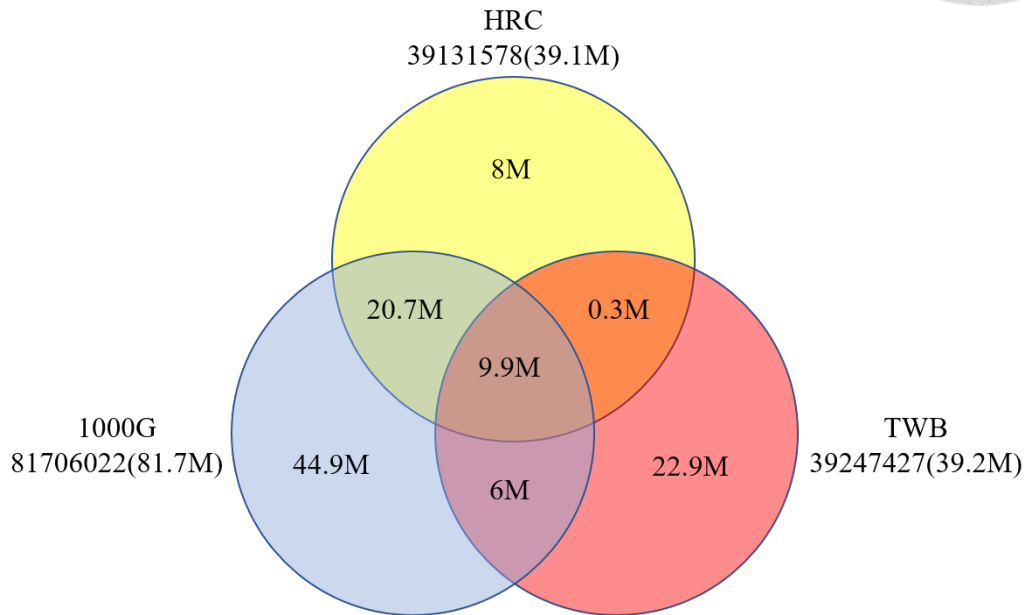


Figure 5 The Venn Diagram of the common reference panels and TWB

Imputation

※File option

Project name

File upload (vcf.gz/bed bim fam) No file chosen

Use URL to upload

Chromosome number

※if want to impute the whole chromosome, please select the function

Impute the full chromosome

Start position

End position

※Quality control option

Minor allele frequency option

Single SNP missing rate option

Individual SNP missing rate option

Hardy-Weinberg Equilibrium option

※Reference option (if need to use custom reference, please select 'Custom')

Reference group select

-----Merge reference-----

※if need to use the merge reference or two phased reference function, please select the second reference that different with the first reference.

Improved reference panels

Merge reference panels

-----Custom reference-----

※Reference option (if need to use custom reference, please upload the reference files)

File upload (.sample, .legend.gz, .hap.gz, genetic_map.txt) No file chosen



Figure 6 The imputation page of the MI-System

MI-System HomePage Function ▾ Result Page Clean id

User Result Page

Please select your project and download here

Imputation

Choose a Project:

Reference


Choose a Project:

Split_chromosome

Choose a Project:

LiftOver

Choose a Project:



All right reserve.

Figure 7 The result page of the MI-System



MI-System HomePage Function ▾ Result Page Tutorial Clean id

Download Page

Function: imputation

[Delete](#)

Project_2021_5_17_5

Date: 2021-05-18 16:27

Chromosome number: 22

Impute the full chromosome: True

Start position: None

End position: None

Minor allele frequency option: --maf 0.05

Single SNP missing rate option: --geno 0.05

Individual SNP missing rate option: --mind 0.05

Hardy-Weinberg Equilibrium option: None

Reference group select: EAS

Progress: Done

※**Result download here**

[PLINK format](#) [Project_2021_5_17_5](#)

[VCF format](#) [Project_2021_5_17_5](#)

※**Log**

[Quality control](#) [Pre-phasing 1](#) [Pre-phasing 2](#) [Pre-phasing 3](#) [Imputation](#)

Figure 8 The download page of the MI-System

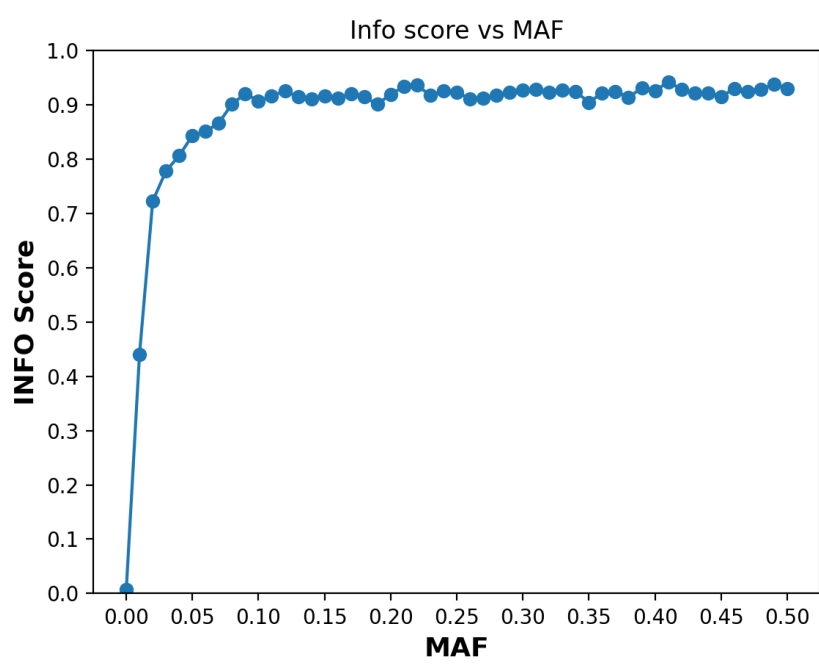


Figure 9 The info score distribution plot of the imputation result

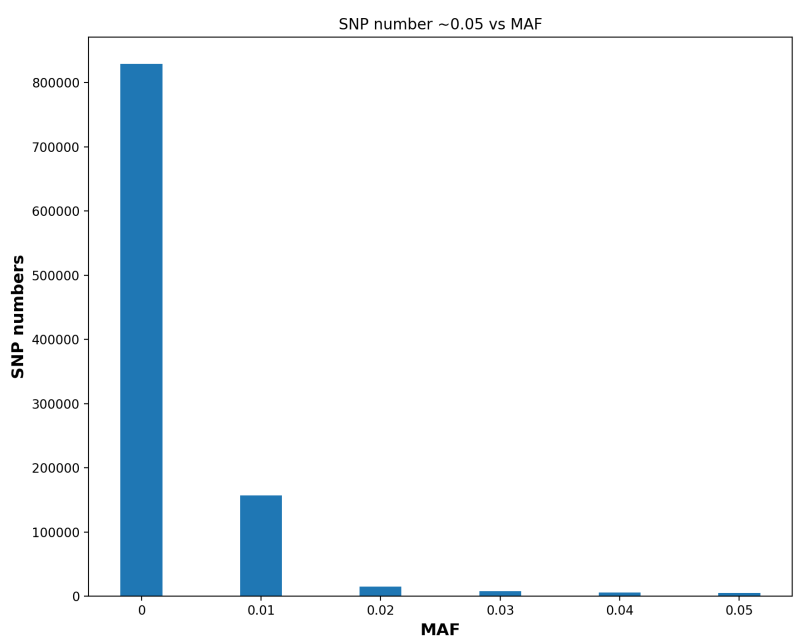


Figure 10 The rare SNPs distribution plot (rare variant) of the imputation result

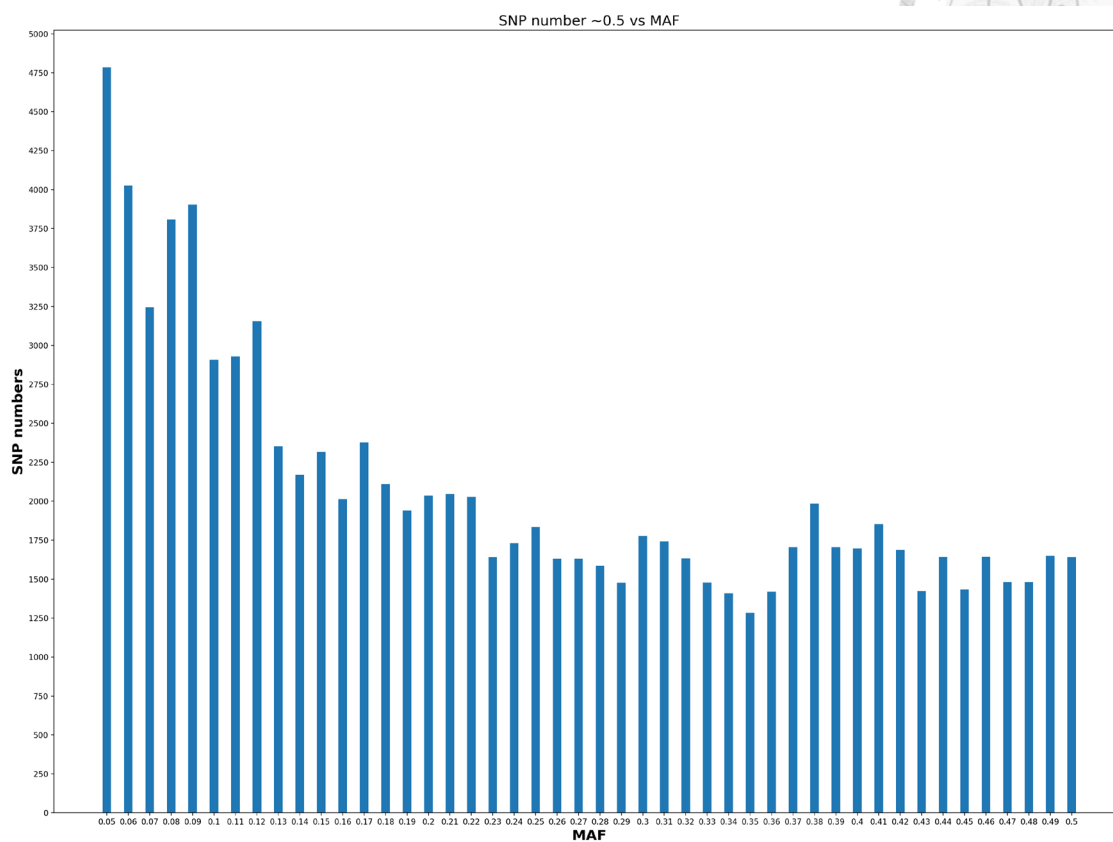


Figure 11 The common SNPs distribution plot (common variant) of the imputation result

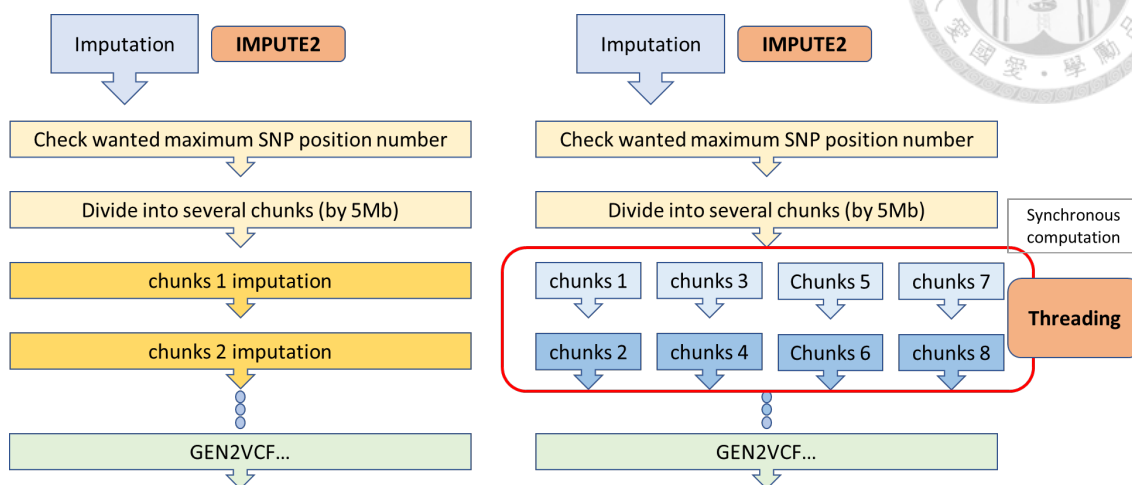


Figure 12 The parallel computation design of the imputation

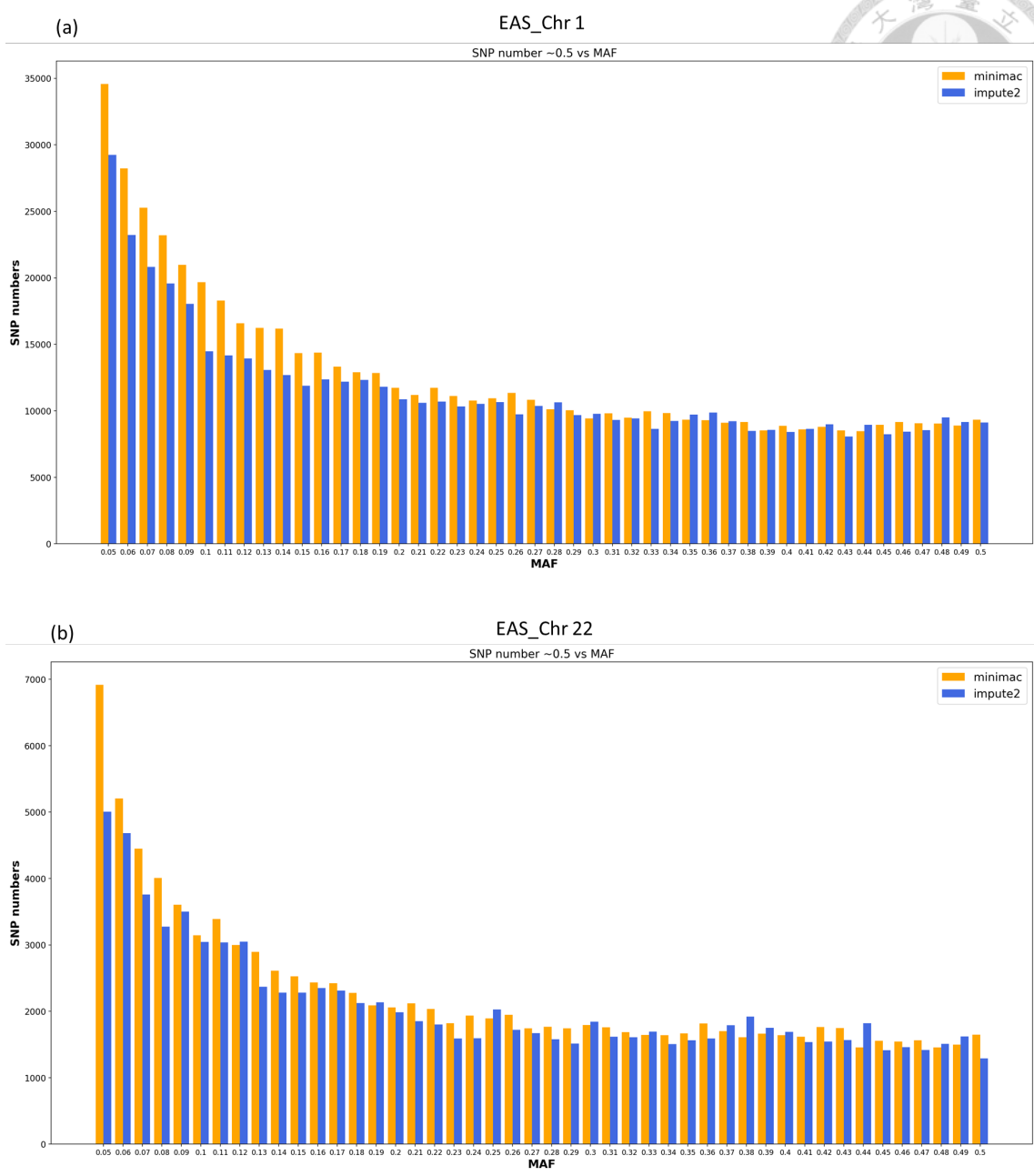
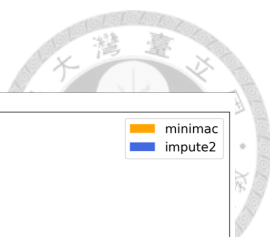


Figure 13 The common SNPs distribution plots from MI-System and Michigan Imputation server (EAS group)

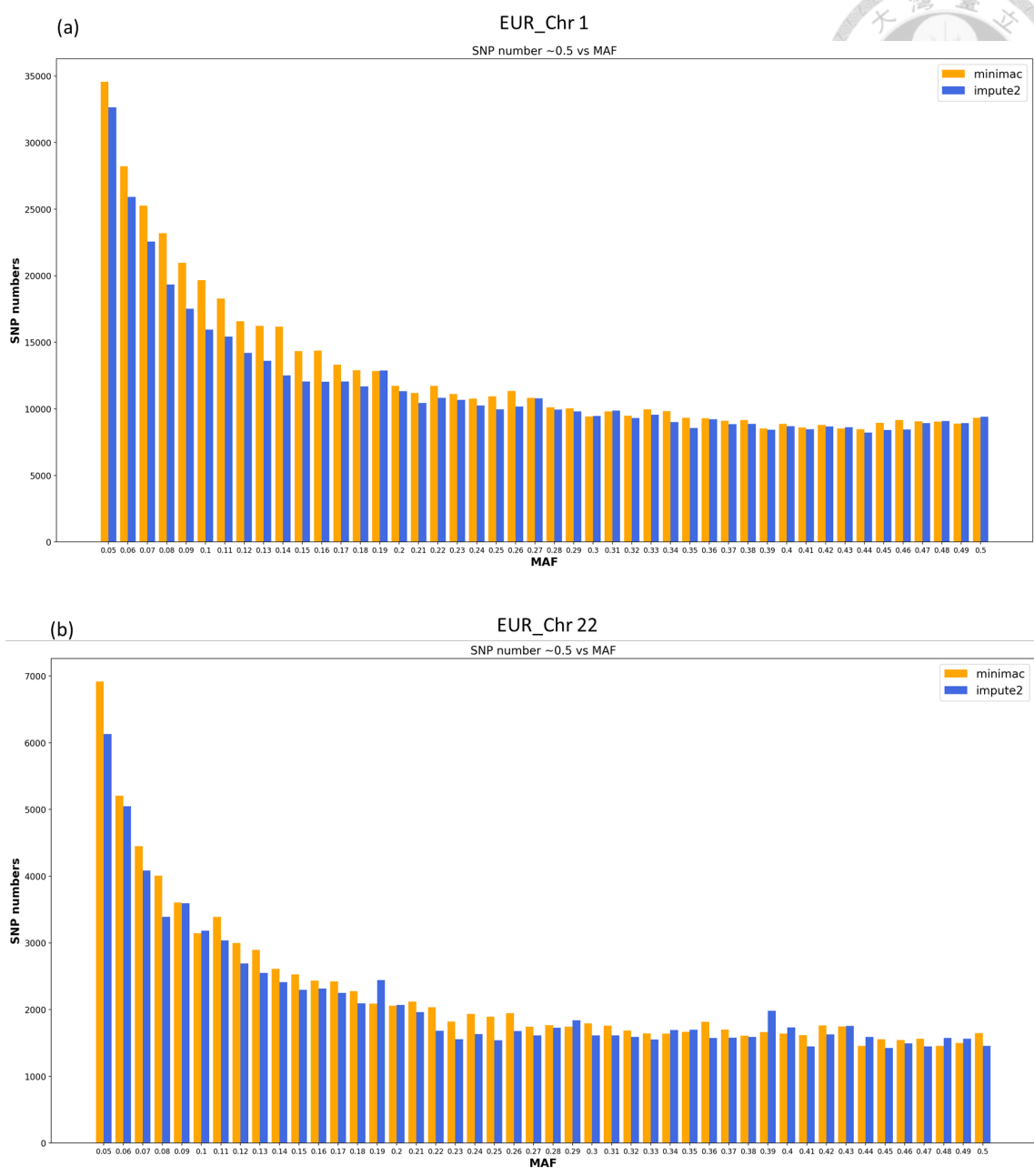
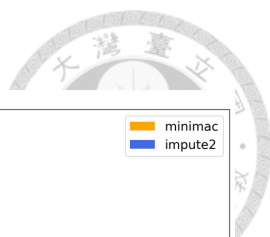


Figure 14 The common SNPs distribution plots from MI-System and Michigan Imputation server (EUR group)

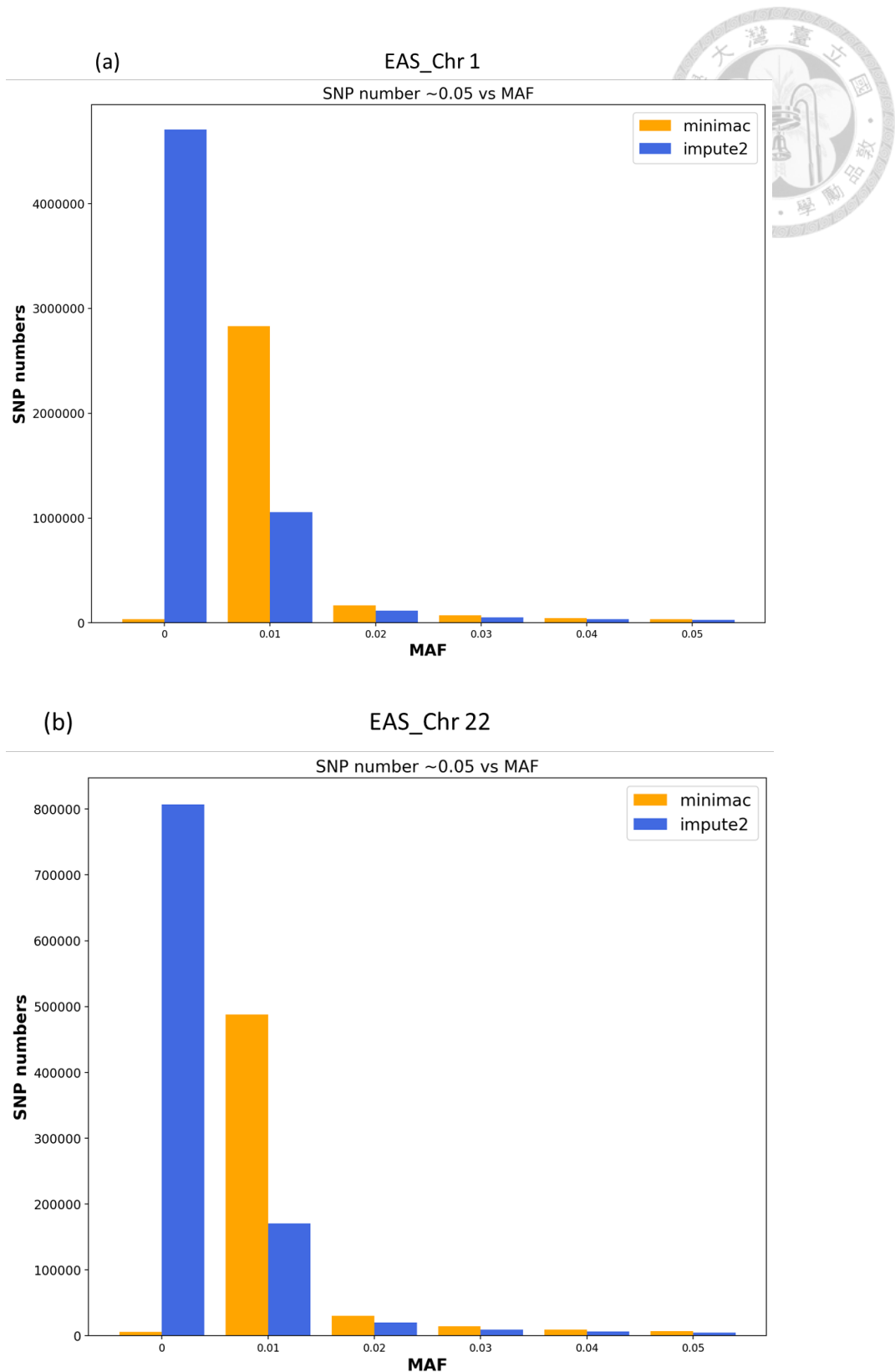


Figure 15 The rare SNPs distribution plots from MI-System and Michigan Imputation server (EAS group)

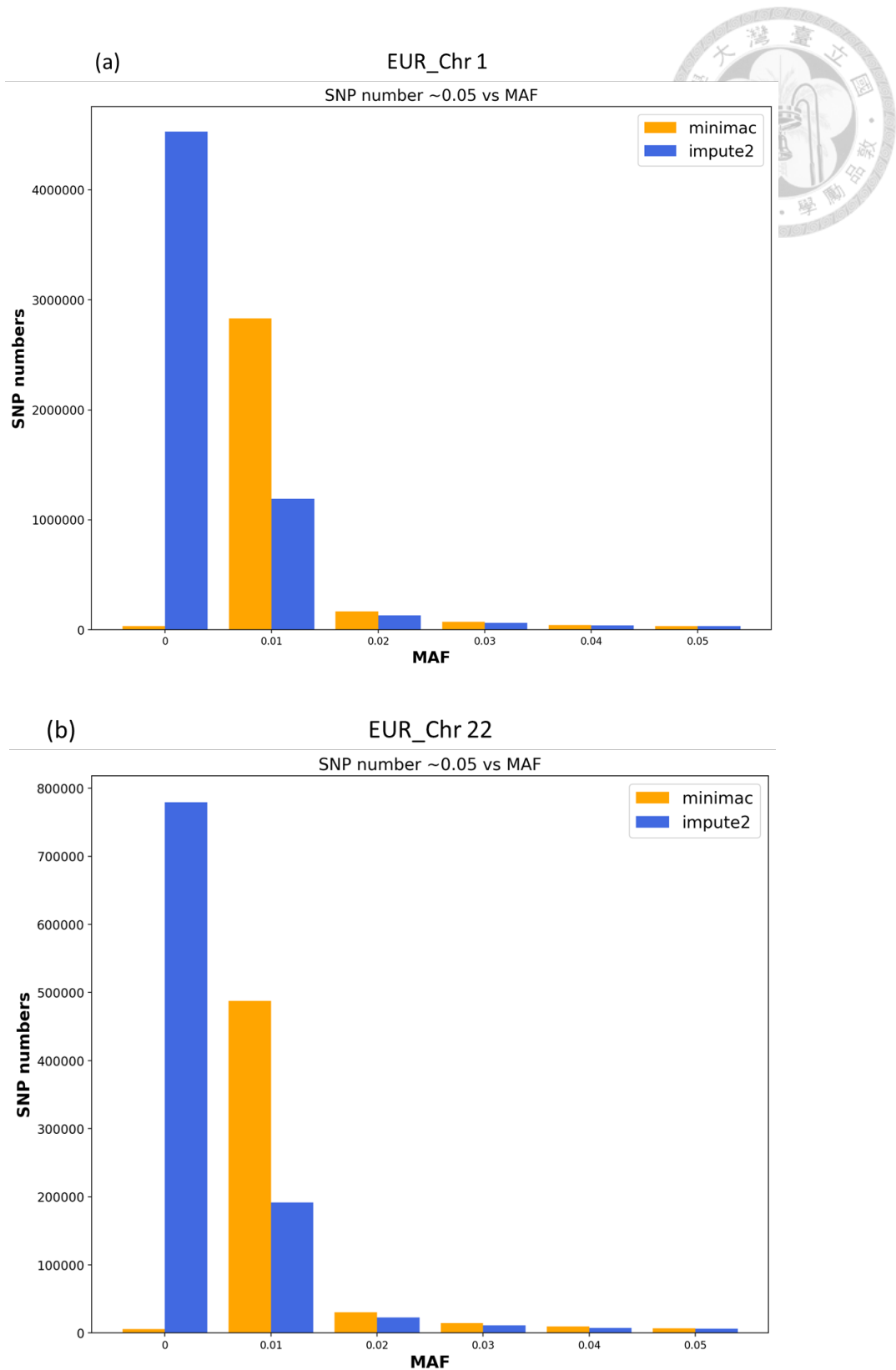


Figure 16 The rare SNPs distribution plots from MI-System and Michigan Imputation server (EAS group)

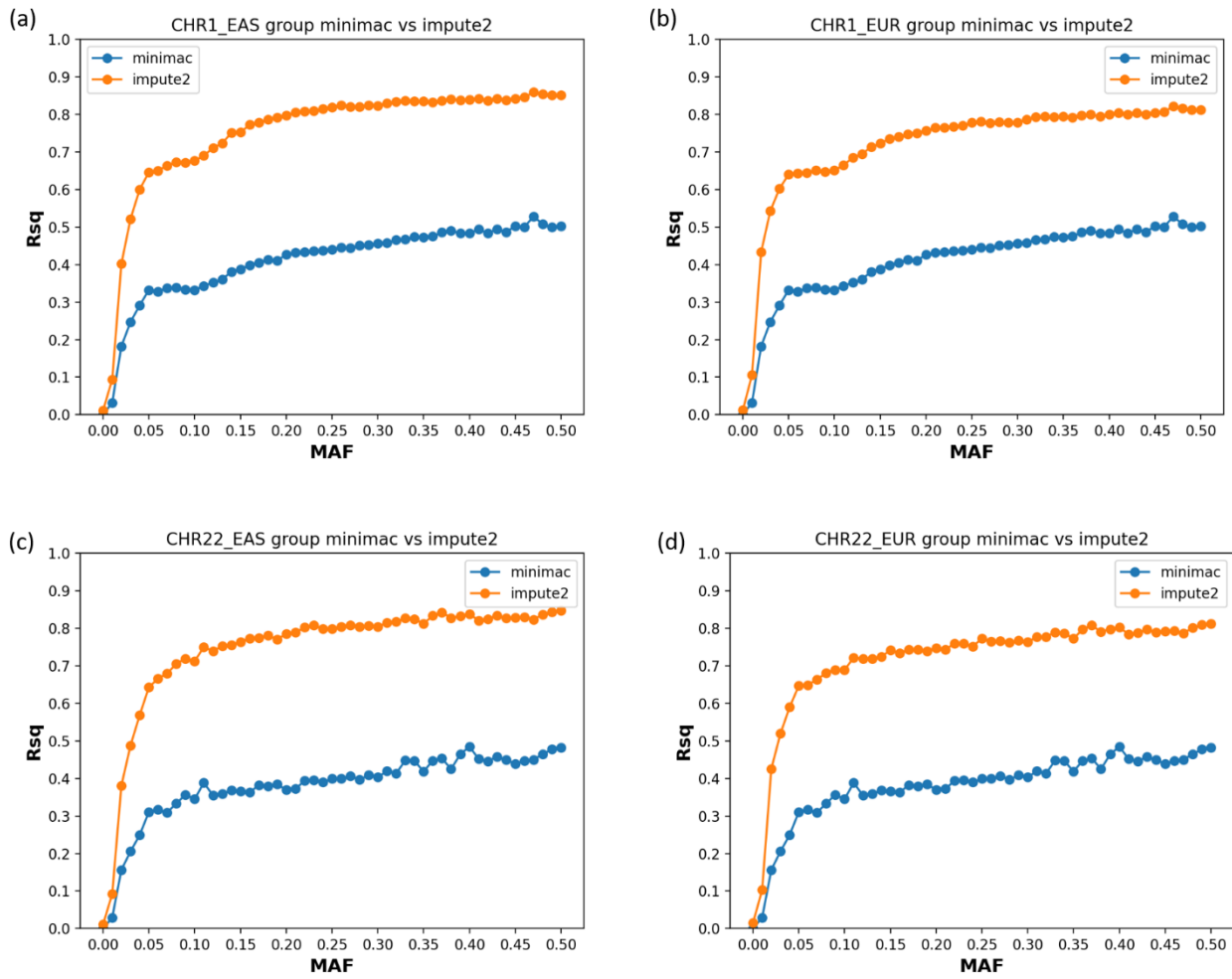
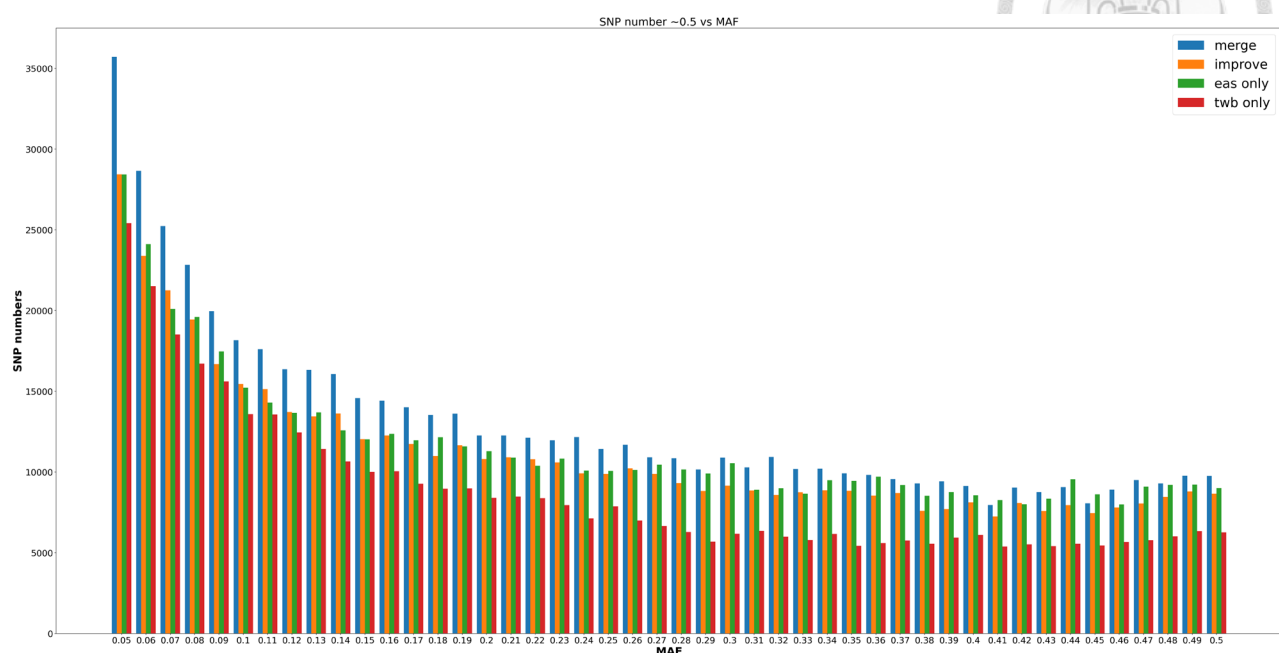


Figure 17. R_{sq} comparison plots of accuracy between Minimac4 and impute2



(a) Chromosome1



(b) Chromosome12

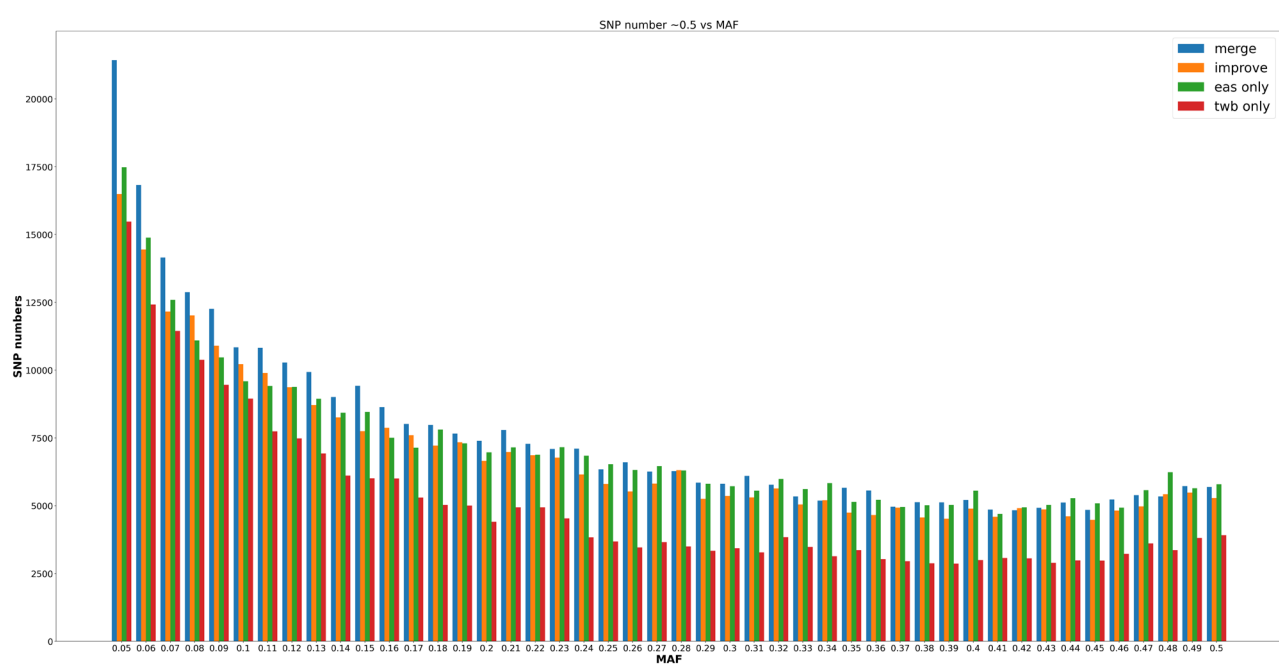


Figure 18. Comparison of common SNPs imputed for 4 different reference panels

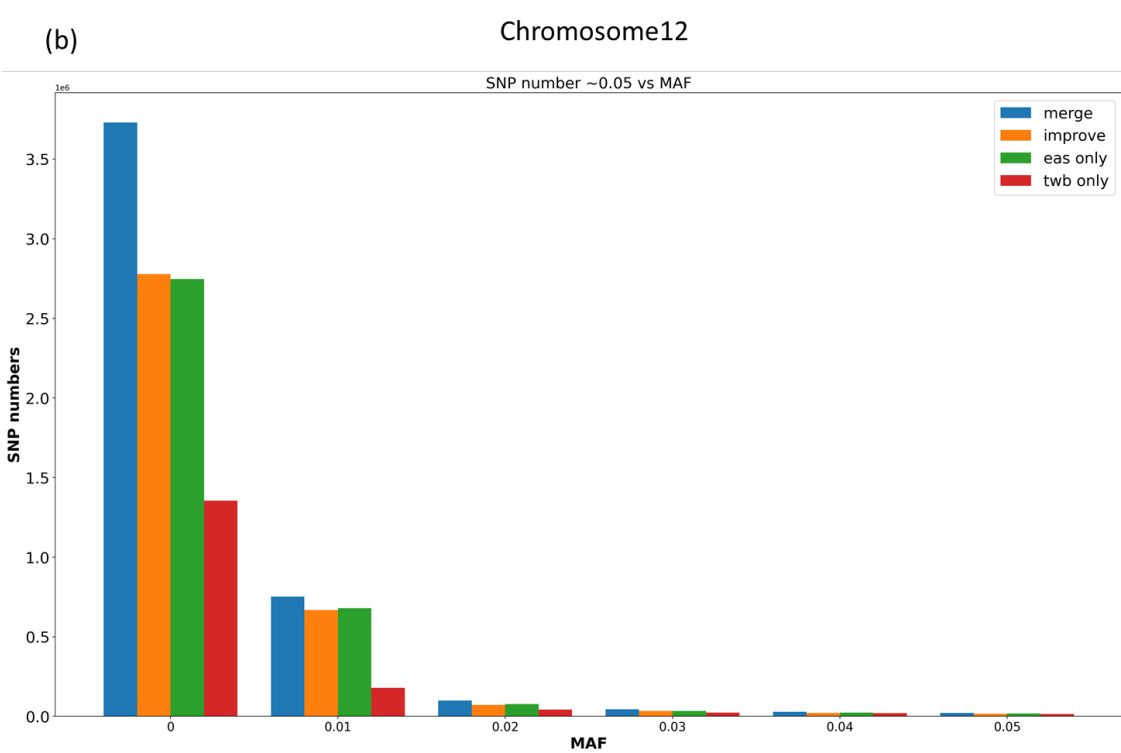
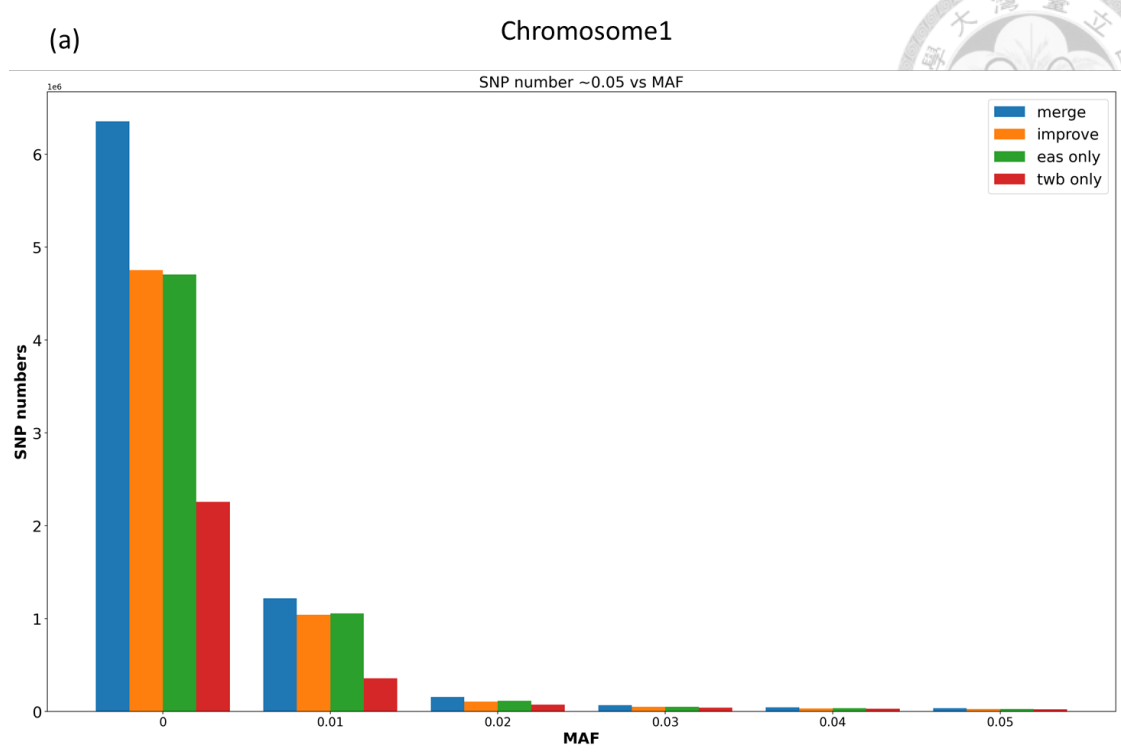
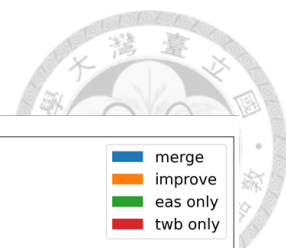


Figure 19. Comparison of rare SNPs imputed for 4 different reference panels

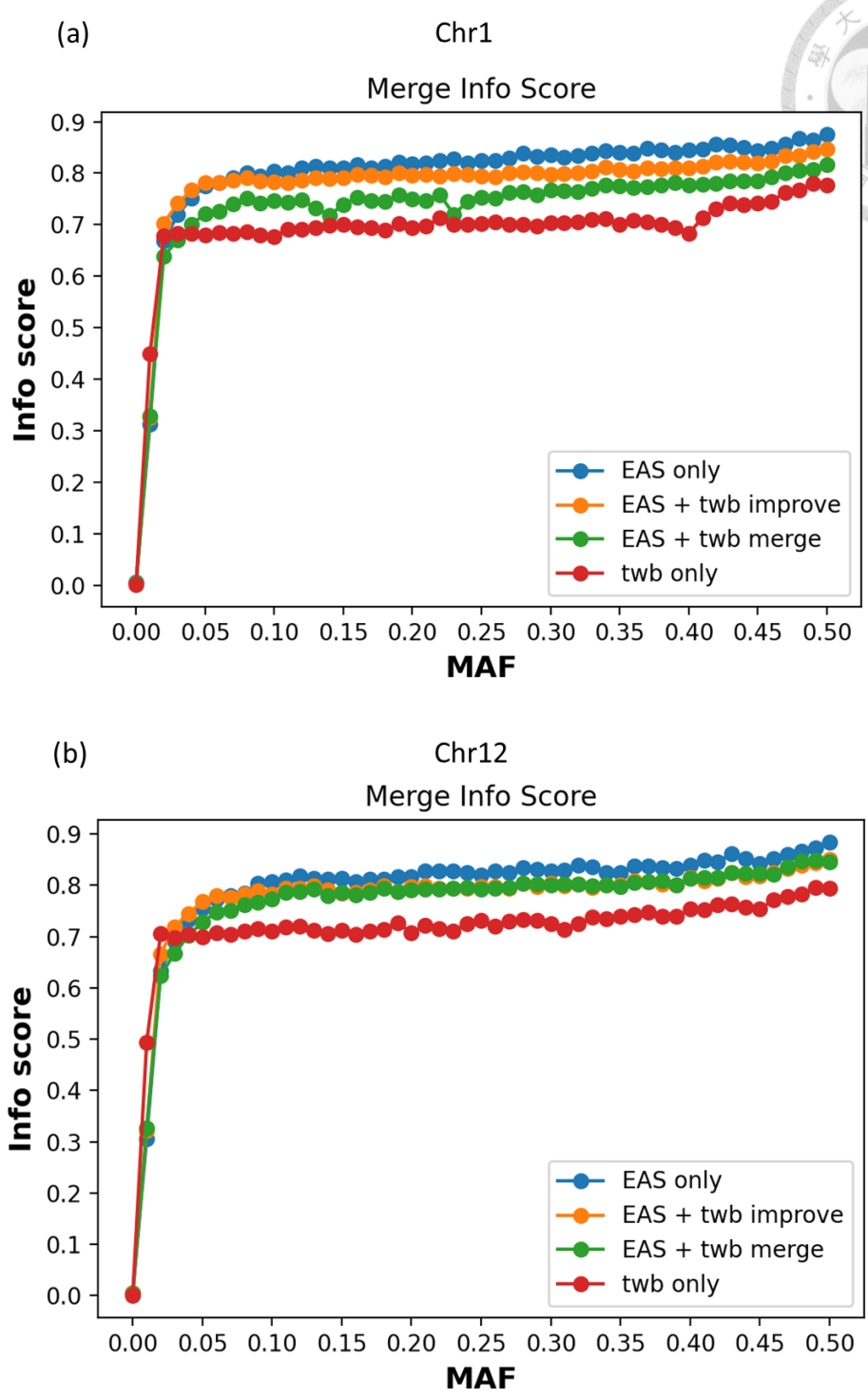


Figure 20. Accuracy comparison between 4 reference panels (Total SNPs)

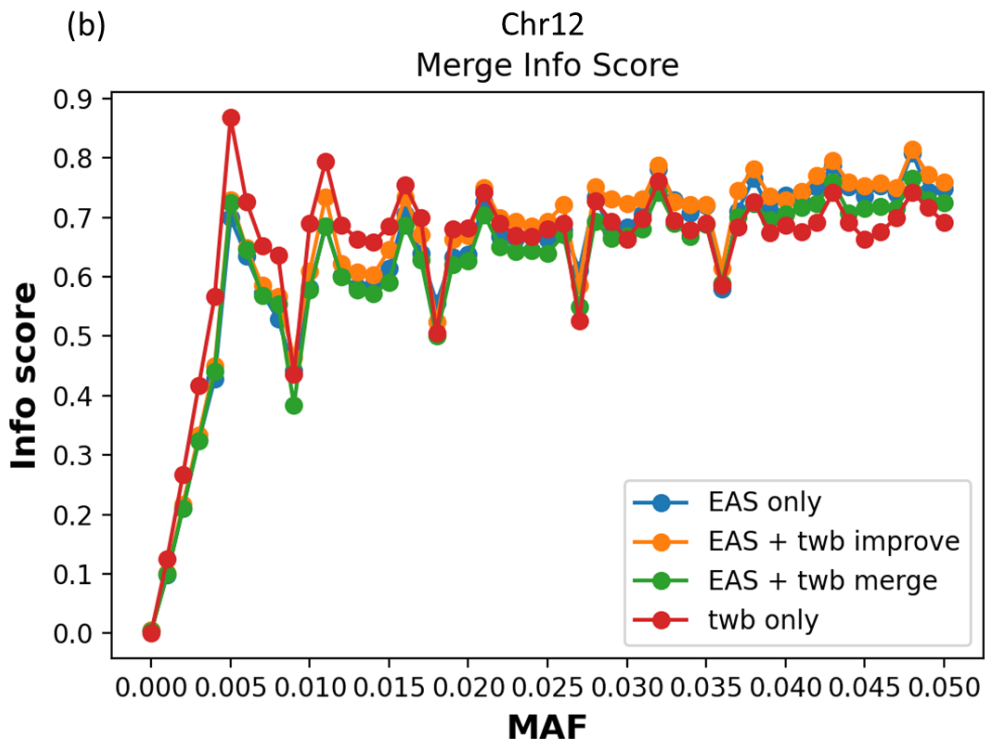
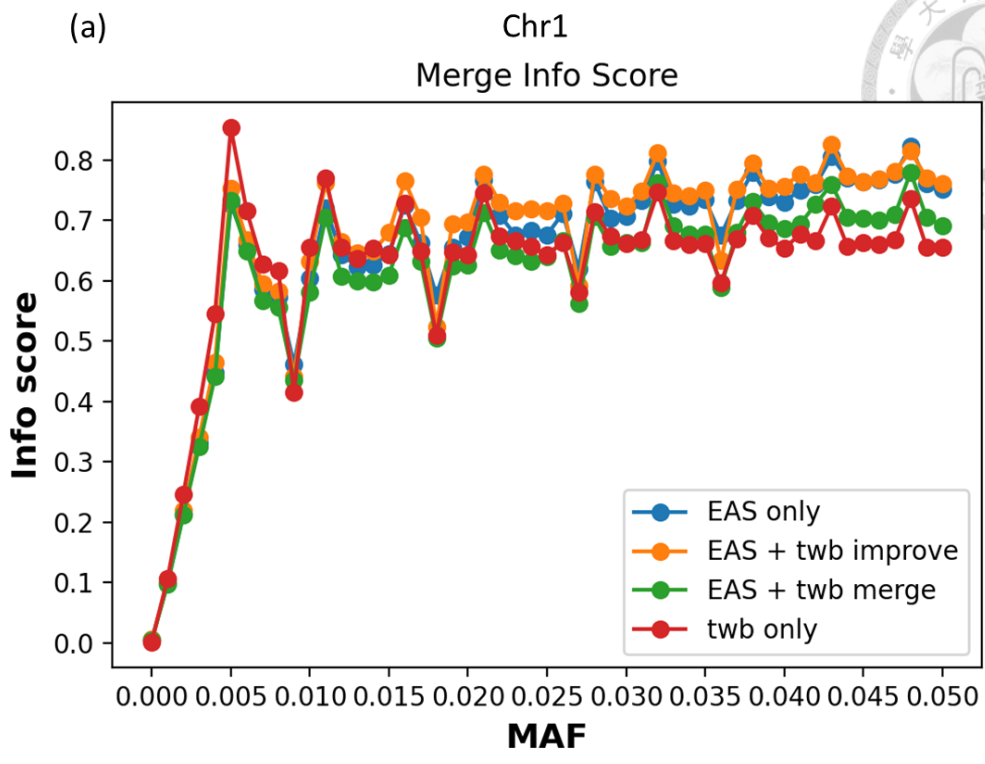


Figure 21. Accuracy comparison between 4 reference panels (Rare SNPs)



MI-System HomePage Function ▾ Result Page Clean id

Split chromosome

※File option

Project name

Data name

File upload (vcf.gz/bed bim fam) No file chosen

Use URL to upload

Allow google drive link

※Notice that the link need to open the authority of the google drive

File upload url (vcf.gz)

File upload url (bed)

File upload url (bim)

File upload url (fam)


 *All right reserve.*

Figure 22 The split chromosome page of the MI-System



LiftOver

※File option

Project name

Data name

Convert function ▾

File upload (vcf.gz/bed bim fam) No file chosen

Use URL to upload

Allow google drive link

※Notice that the link need to open the authority of the google drive

File upload url (vcf.gz)

File upload url (bed)

File upload url (bim)

File upload url (fam)



All right reserve.

Figure 23 The liftover page of the MI-System



Create Reference Panel

※File option

Project name

File upload (vcf.gz/bed bim fam) No file chosen

Chromosome number ▾

Use URL to upload

Allow google drive link

※Notice that the link need to open the authority of the google drive

File upload url (vcf.gz)

File upload url (bed)

File upload url (bim)

File upload url (fam)



All right reserve.

Figure 24 The Create reference page of the MI-System

Table:



Validation Imputation accuracy

SNP ID	REF	ALT	Total	Missing	Error	Accuracy
rs4822389	T	C	188	1	1	98.9%
rs9609329	A	C	188	1	1	98.9%
rs135020	G	A	188	0	0	100%
rs627396	C	T	188	1	1	98.9%
rs6008294	T	C	188	0	0	100%

Table 1 The validation of imputation accuracy



	CGM server (origin)	CGM server (Multi-threading)
CHR22 – 188 sample 6404 SNPs (1000G-EAS)	2hr and 20min	30 min
CHR22- 304 sample 7294 SNPs (1000G-EAS)	3hr and 16hr	44 min
CHR1 – 455sample 41592 SNPs(1000G-EAS)	28 hr	2hr and 40 min (20 threads)

Table 2 The imputation cost time of the IMPUTE2



Server comparison

	MI-system	Michigan imputation server	Sanger imputation service
<i>Imputation tools</i>			
Quality Control	PLINK	PLINK	—
Phasing	SHAPEIT2	EAGLE2	EAGLE2/SHAPEIT2
Imputation	IMPUTE2	Minimac4	PBWT
<i>Reference panels</i>			
1000 Genomes	v	v	v
HRC		v	
GAsP		v	
CAAPA		v	
HapMap2		v	
HapMap3	v		
UK10K			v
AGR			v
TWB	v		
Custom	v		
<i>Support functions</i>			
Split chromosome	v		
LiftOver	v		
Make reference panels	v		
Batch upload		v	v
<i>System</i>			
Support platforms	windows/mac	windows/mac	windows/mac

Note. HRC = Haplotype Reference Consortium, GAsP = Genome Asia Pilot, CAAPA = Consortium on Asthma among African-ancestry Populations in the Americas, AGR = African Genome Resources, TWB =Taiwan Biobank

Table 3 The comparison table between several imputation servers and MI-System



Supplementary figure:

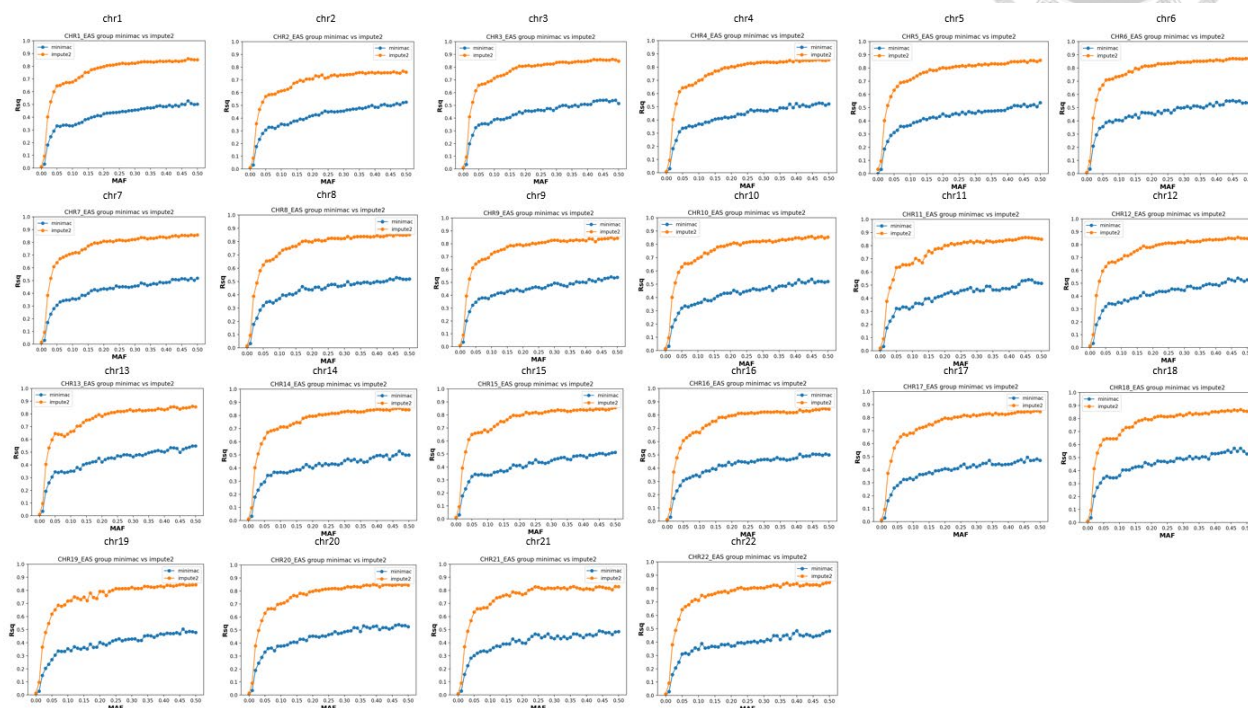


Figure S1. Rsq comparison plots of accuracy between Minimac4 and impute2 in all chromosomes (EAS group)

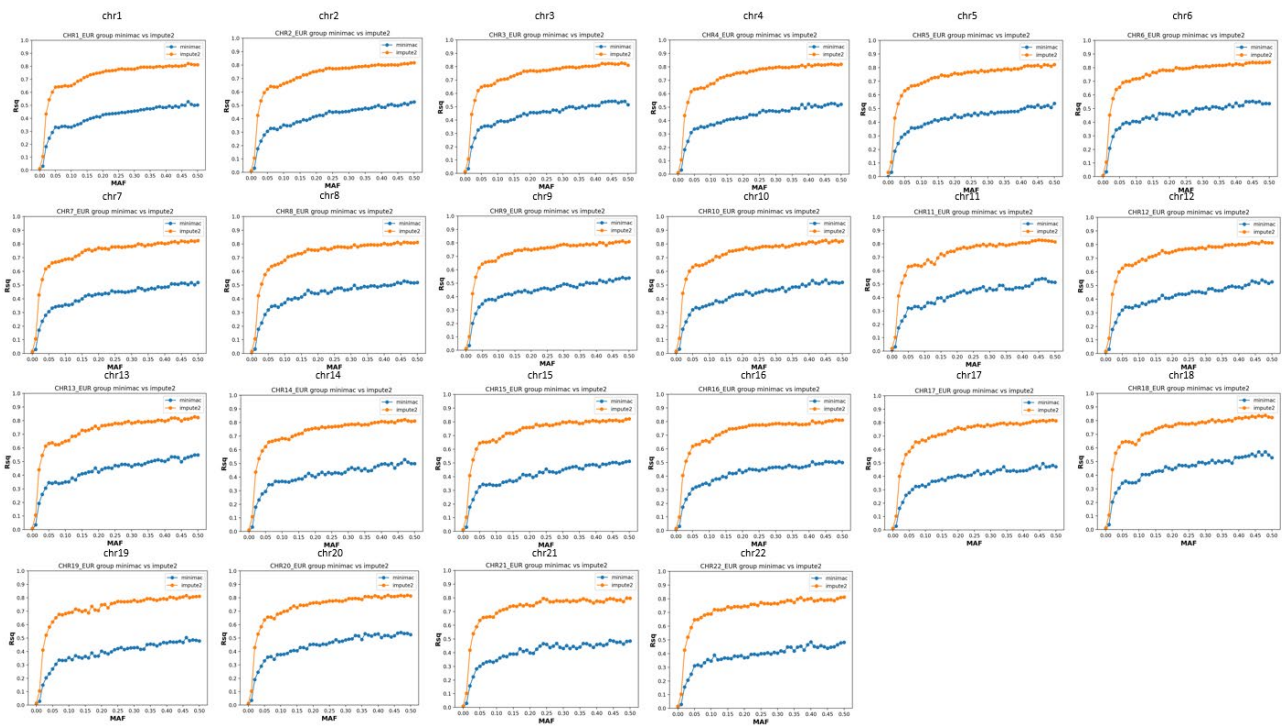


Figure S2. Rsq comparison plots of accuracy between Minimac4 and impute2 in all chromosomes (EUR group)

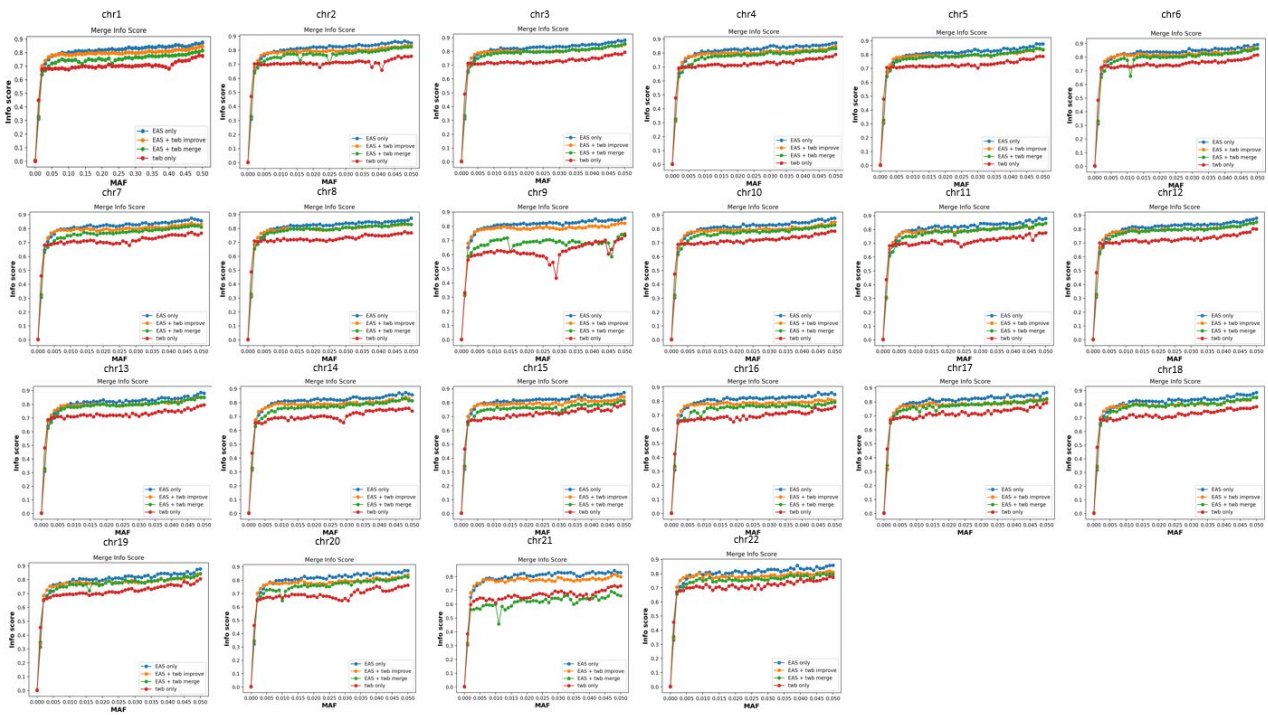


Figure S3. Accuracy comparison between 4 reference panels in all chromosomes (Total SNPs)

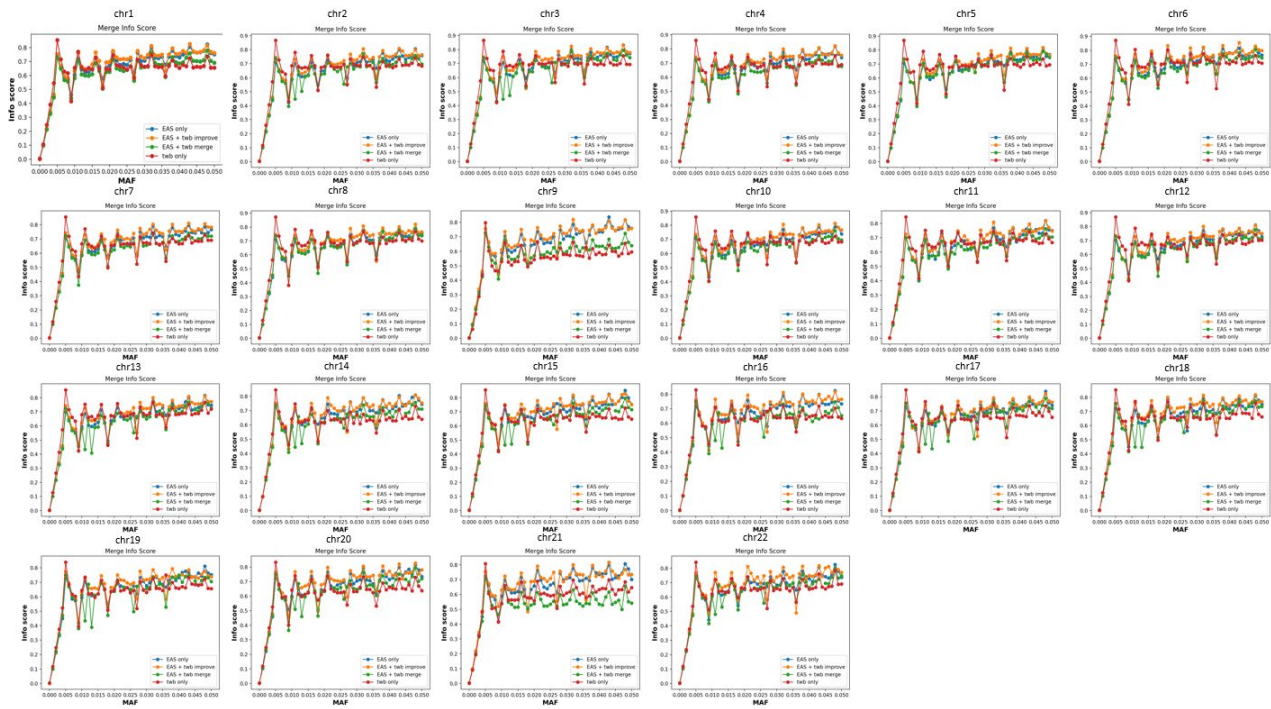


Figure S4. Accuracy comparison between 4 reference panels in all chromosomes (Rare SNPs)